Coursera Statistical Inference Project 1

Seth Dobrin

February 8, 2016

Title (give an appropriate title) and Author Name Overview: In a few (2-3) sentences explain what is going to be reported on. Simulations: Include English explanations of the simulations you ran, with the accompanying R code. Your explanations should make clear what the R code accomplishes. Sample Mean versus Theoretical Mean: Include figures with titles. In the figures, highlight the means you are comparing. Include text that explains the figures and what is shown on them, and provides appropriate numbers. Sample Variance versus Theoretical Variance: Include figures (output from R) with titles. Highlight the variances you are comparing. Include text that explains your understanding of the differences of the variances. Distribution: Via figures and text, explain how one can tell the distribution is approximately normal.

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. You should

Show the sample mean and compare it to the theoretical mean of the distribution. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution. Show that the distribution is approximately normal. A sample set of headings that could be used to guide the creation of your report might be:

```
library(knitr)
library(ggplot2)
```

knitr settings

```
knitr::opts_chunk$set(echo = TRUE, tidy = TRUE, fig.width = 8, fig.height = 5,
    cache = TRUE)
```

Use set.seed so that simulation can be repreated. Since random numbers in R arent really ramdom seeting the seed provides the same value to the random number generating function.

```
set.seed(1)
```

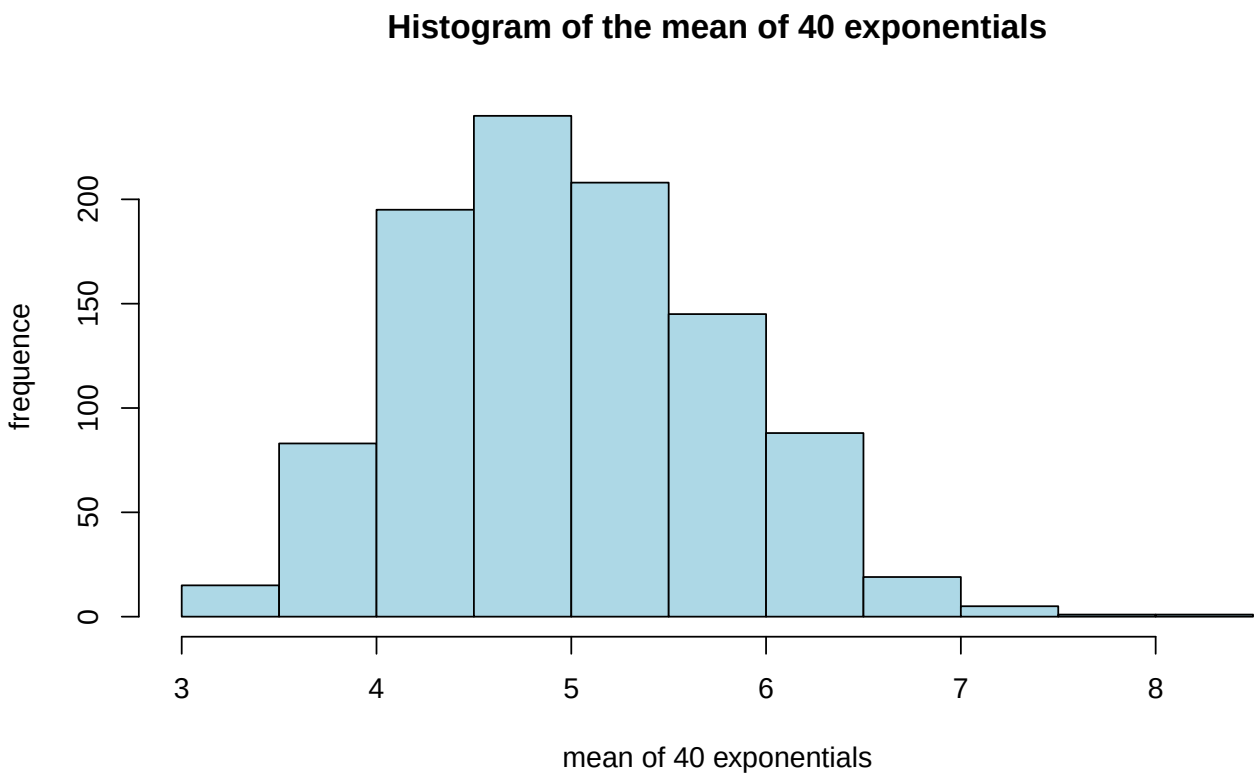
Simulation

The exponential distribution is simulated in R using rexp(n, lambda). Using this command n is the number of simulaitons and lambda is the rate parameter. Both the mean of exponential distribution and the standard deviation are 1/lambda. For all simulations lambda = 0.2.

```
lambda <- 0.2
sampleSize <- 40
simulations <- 1000
simMatrix <- matrix(rexp(simulations * sampleSize, rate = lambda), simulations,
    sampleSize)
```

The simulation mean is generated and plotted below.

```
simMean <- rowMeans(simMatrix)
hist(simMean, xlab = "mean of 40 exponentials", ylab = "frequence", main = "Histogram of the mean of 40
    col = "lightblue")
```



Calculate the sample mean vs. theoretical mean

```
sampleMean <- mean(simMean)
cat("sample mean: ", sampleMean)
```

```
## sample mean:  4.990025
```

The sample mean approximates the theoretical mean.

```
theoreticalMean <- 1/lambda
cat("theoretical mean: ", theoreticalMean)
```

```
## theoretical mean:  5
```

Sample variance vs. theoretical variance

```r
sampleVariance <- var(simMean)
cat("sample variance: ", sampleVariance)
```

```
## sample variance:  0.6177072
```

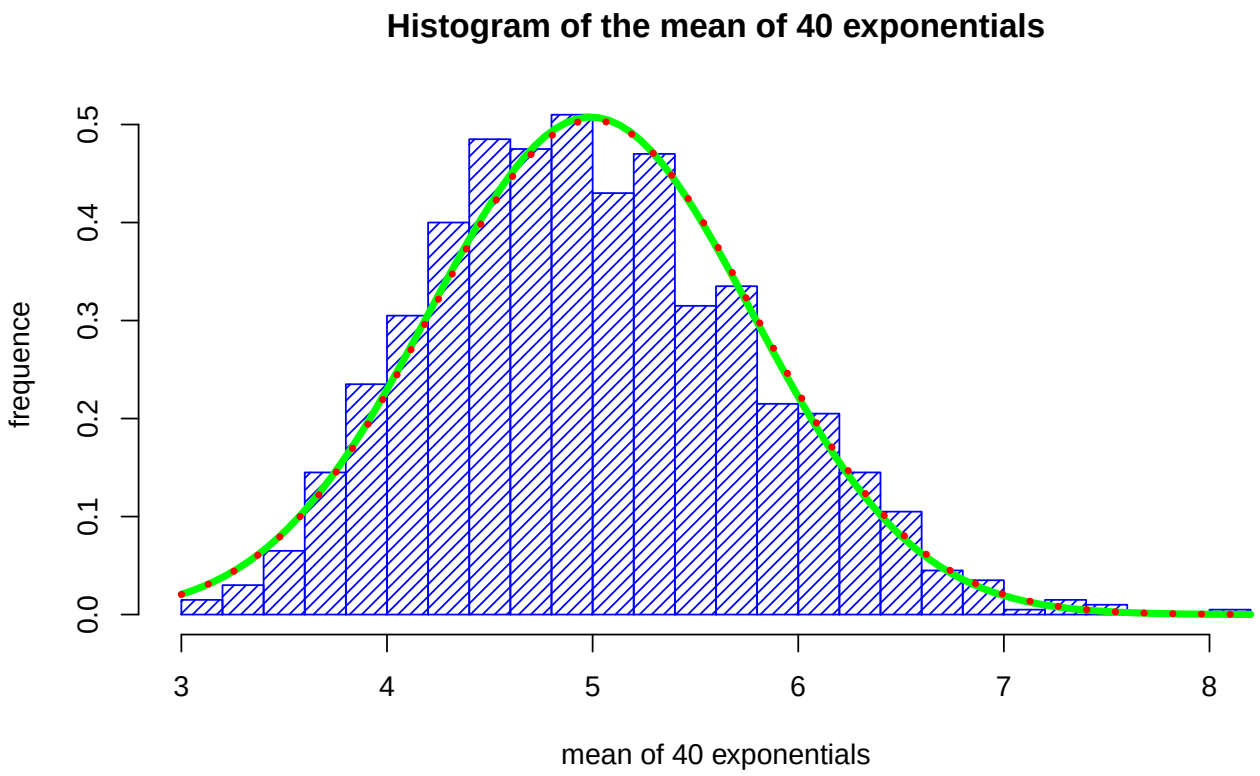
```r
theoreticalVariance <- (1/lambda)^2/sampleSize
cat("theoretical variance: ", theoreticalVariance)
```

```
## theoretical variance:  0.625
```

Distribution

Create a histogram of the mean of 40 exponentials (blue hashed bars) with the sample mean (green line) and theoretical means (dotted red line) plotted

```r
hist(simMean, density = 20, breaks = 20, prob = TRUE, xlab = "mean of 40 exponentials",
    ylab = "frequence", main = "Histogram of the mean of 40 exponentials", col = "blue")
curve(dnorm(x, mean = sampleMean, sd = sqrt(sampleVariance)), col = "green",
    lwd = 4, add = TRUE, yaxt = "n")
curve(dnorm(x, mean = theoreticalMean, sd = sqrt(theoreticalVariance)), col = "red",
    lwd = 4, lty = "dotted", add = TRUE, yaxt = "n")
```



This plot shows that the mean of 40 exponentials can can be approximated with the normal distribution.

Additionally, the mean and variance of sample data approximate a normal distribution. The confidence intervals are as follows:

```
confidenceSample <- round(mean(simMean) + c(-1, 1) * 1.96 * sd(simMean)/sqrt(sampleSize),
    3)
cat("sample confidence interval: ", confidenceSample)
```

```
## sample confidence interval:  4.746 5.234
```

```
confidenceTheoretical <- theoreticalMean + c(-1, 1) * 1.96 * sqrt(theoreticalVariance)/sqrt(sampleSize)
cat("theoretical confidence interval: ", confidenceTheoretical)
```

```
## theoretical confidence interval:  4.755 5.245
```

A quantile-qulatile or Q-Q plot compares the theroretically expected value to the sample values. If the plotted points approximate the line of slope 1, then the sample distribution is approximately normal. Since the plot below of my theoretical and sample means approximate a line of slope 1 the sample is normally distributed.

```
qqnorm(simMean, main = "Q-Q Plot", xlab = "Theoretical Quantiles", ylab = "Sample Quantiles")
qqline(simMean, col = "blue")
```