Statistical Inference Course Project Part 1 - A Simulation Exercise

Randall Helms

24 October 2016

```
knitr::opts_chunk$set(echo = TRUE)
```

Introduction

This document covers the three questions on the first part of the Statistical Inference course project. This course is part of the Data Science Specialization offered by Johns Hopkins University via Coursera.

This project involves using R to investigate the exponential distribution and then compare it to the Central Limit Theorem.

In this project we have to illustrate and explain the properties of the distribution of the mean of 40 exponentials, involving the following three tasks:

```
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 theoretica
l variance of the distribution.
3. Show that the distribution is approximately normal.
```

In order to do this, the course authors have provided the following parameters to work with:

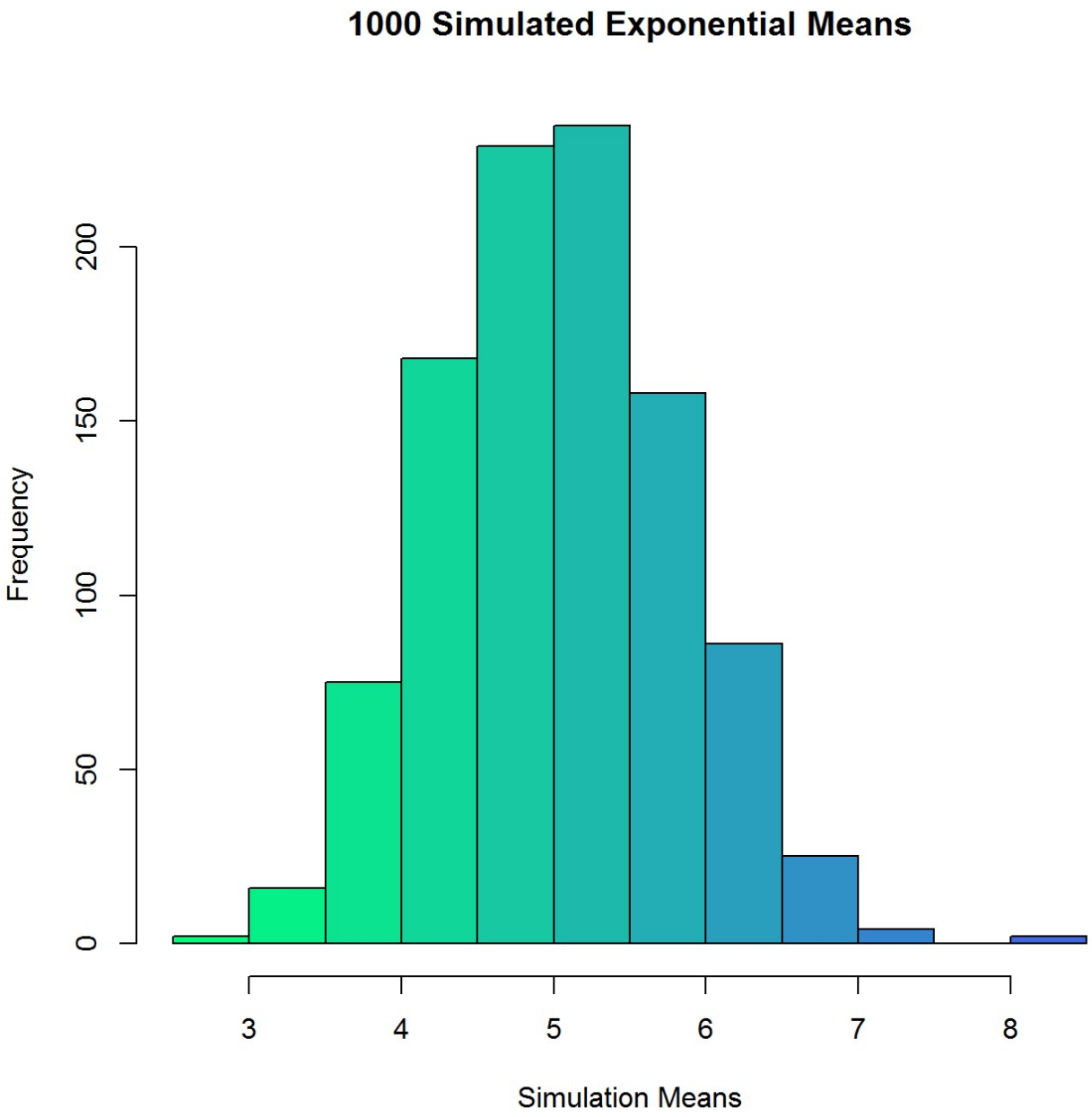
```
set.seed(1980)
lambda = 0.2
n = 40
simulations = 1000

exponential_distribution_simulation <- rexp(n,lambda)

mean_exponential_distribution = 1/lambda

standard_deviation = 1/lambda
```

Building on this, let's simulate the exponentials and then calculate the mean of the simulation:

```
#simulate the exponentials

sim_exp <- replicate(simulations,rexp(n,lambda))

#calculate the mean of exponentials

mean_exp <- apply(sim_exp,2,mean)
```

Let's take things one step further and use that information to plot out a simple histogram showing the distribution of means:

```
library(grDevices)
colfunc<-colorRampPalette(c("springgreen","royalblue")) #create a gradient color,
just because it looks nicer than the block colors
hist(mean_exp,main = "1000 Simulated Exponential Means",xlab="Simulation Means",co
l=colfunc(12))
```



Question 1

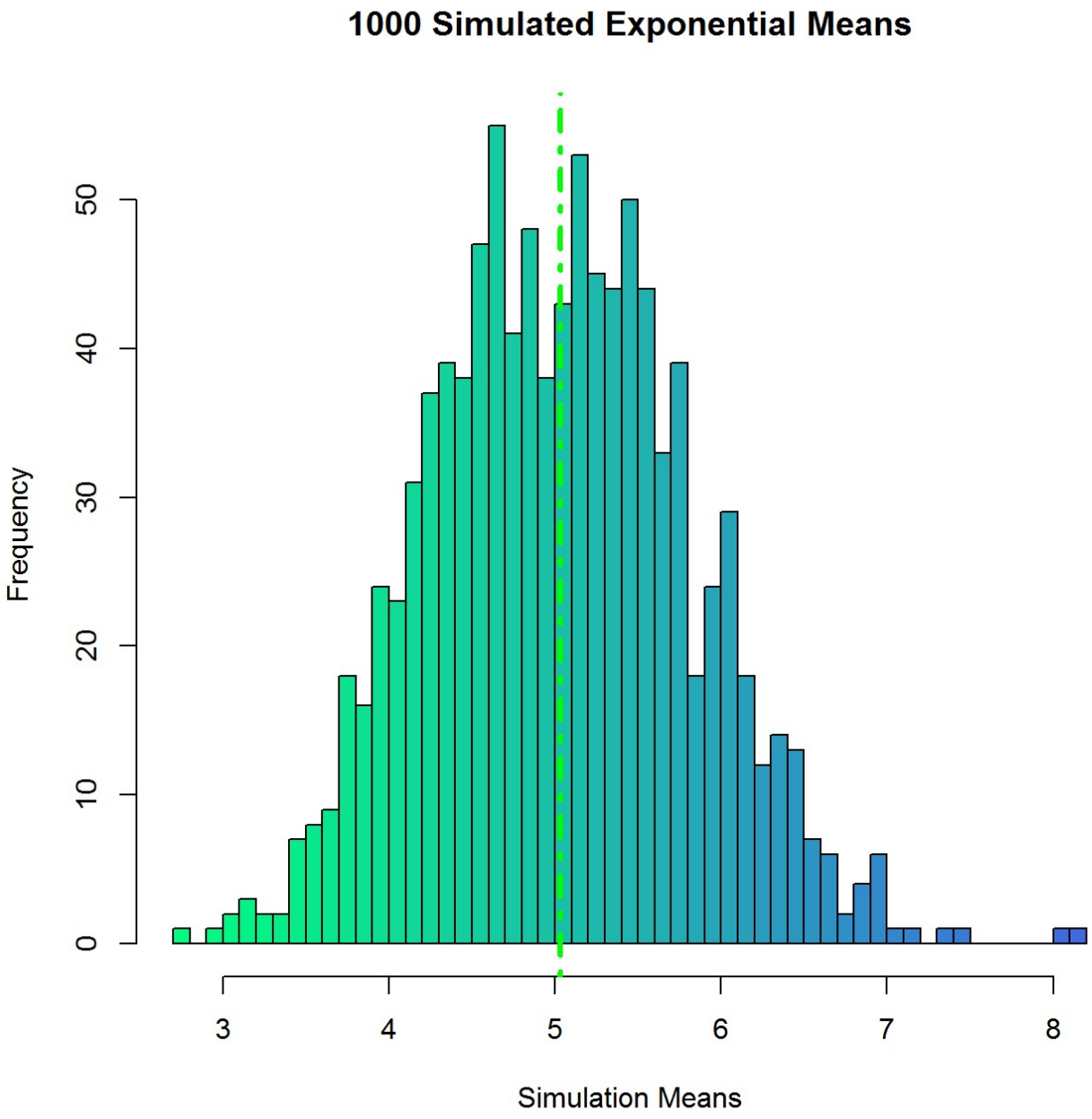
The first question requires us to show the sample mean and compare it to the theoretical mean of the distribution.

Here's how we calculate the sample mean:

```
sample_mean = mean(mean_exp)

#result is 5.030016
```

To compare this to the theoretical mean of the distribution, we can put it on the chart we made before:

```
hist(mean_exp,main = "1000 Simulated Exponential Means",xlab="Simulation Means",co
l=colfunc(55),breaks=40)
abline(v=sample_mean,lwd="3",col="green",lty=4)
```



Since the sample mean is 5.0030016, this is extremely close to the theoretical variance of the distribution, which is 5 (from 1/lambda, with lambda being 0.2).

Question 2

For the second question, we are looking using variance to check how variable the sample is, and then comparing that information to the theoretical variance of the distribution.

To do this successfully, we need to calculate the standard deviations and variances for both the sample and the theoretical distributions:

```
sim_sd <- sd(mean_exp)
sim_var <- sim_sd^2

theory_sd <- (1/lambda)/sqrt(n)
theory_var <- theory_sd^2
```

Let's compare the results:

```
print(paste("Theoretical standard deviation: ",theory_sd, "vs. Simulated standard
deviation: ",sim_sd))
```

[1] "Theoretical standard deviation: 0.790569415042095 vs. Simulated standard deviation: 0.788572277744267"

```
print(paste("Theoretical variance: ",theory_var, "vs. Simulated variance: ",sim_va
r))
```

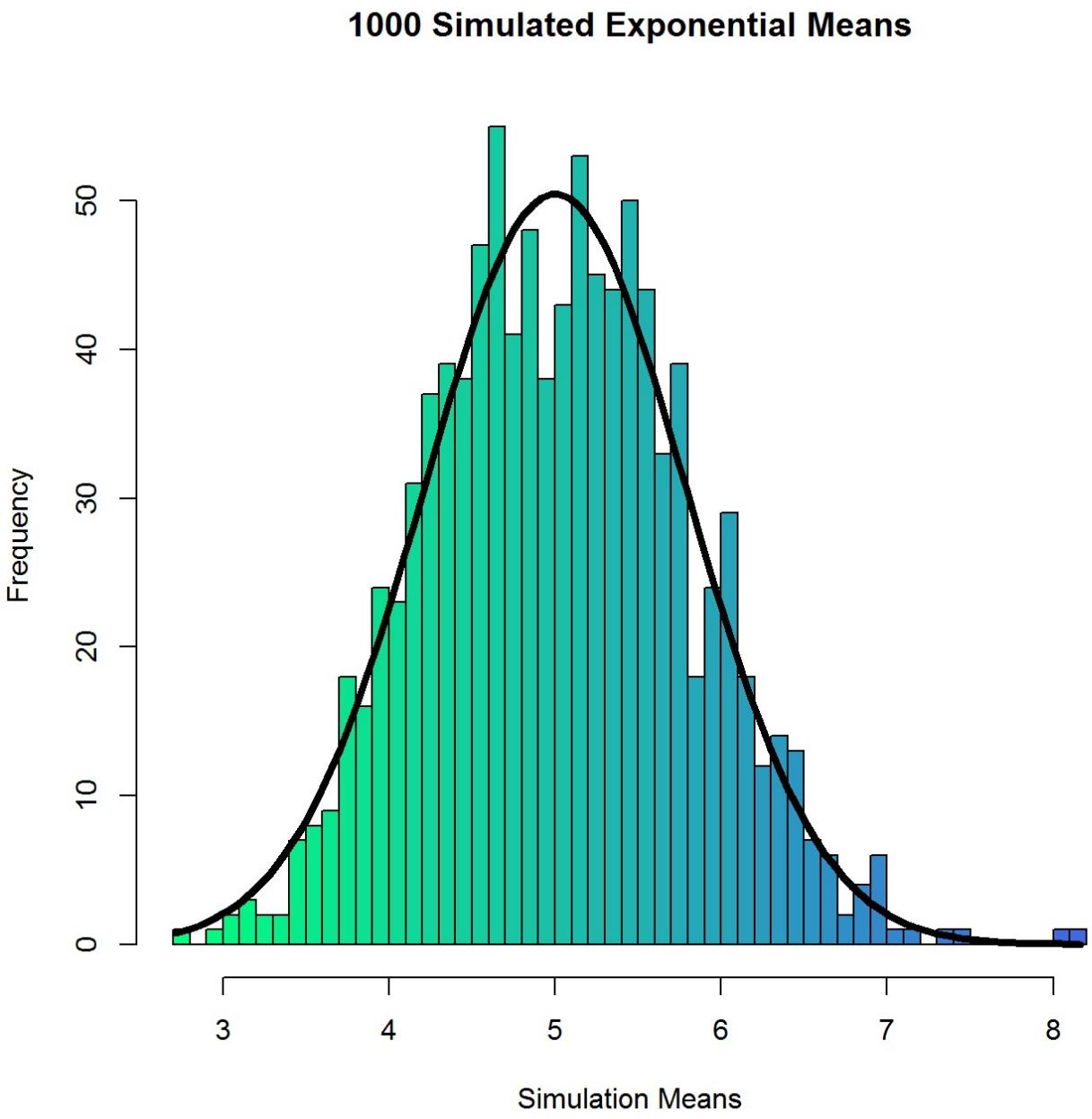
[1] "Theoretical variance: 0.625 vs. Simulated variance: 0.621846237226781"

In both of these cases, the theoretical values are very close to the actual sample values.

Question 3

The last question asks us to show that the distribution is approximately normal; the best way to do so is to re-draw a histogram of the distribution and then fit a line over that to show the distibution:

```
hist(mean_exp,main = "1000 Simulated Exponential Means",xlab="Simulation Means",co
l=colfunc(55),breaks=40)

x <- seq(min(mean_exp),max(mean_exp),length=100)
y <- dnorm(x, mean = 1/lambda, sd = theory_sd)

lines(x,y*100,col="black",lty=1,lwd=4)
```



As you can see once the lines are placed over the histogram, the sampled data is approximately normally distributed.