Coursera Statistical Inference Peer Assignment Part I

Carlos Astrada

10/28/2016

Overview

In this project I 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. I will set lambda = 0.2 for all of the simulations. I will investigate the distribution of averages of 40 exponentials.

In this simulation, I will illustrate the properties of the distribution of the mean of 40 exponentials. Our goals are to:

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.

```r
# Set our working directory
setwd("/Users/charlyastrada/Sites/dataanalysis/stat_inference")

# Load libraries
library(ggplot2)

# Set seed for reproducibility
set.seed(2016)
```

Simulations

First, we will set the variables we will use to run our simulation.

```r
n <- 40
lambda <- 0.2
sims <- 5000
```

Second, we will run the simulation for an exponential distribution 5000 times and store into a matrix.

```r
exp.dist.sim <- matrix(rexp(n * sims, rate = lambda), sims)
```

Third, we will take the means of each simulation from `exp.dist.sim` using apply for each row.

```r
exp.dist.sim.means <- apply(exp.dist.sim, 1, mean)
```

Sample Mean vs Theoretical Mean

The theoretical mean for an exponential distribution is `1/lambda` or, if we use the value of 0.2 as we do in our simulation, we should see a mean of `1/0.2` , or 5. Let's take the mean of our simulation data to compare to the theoretical mean of: 5.

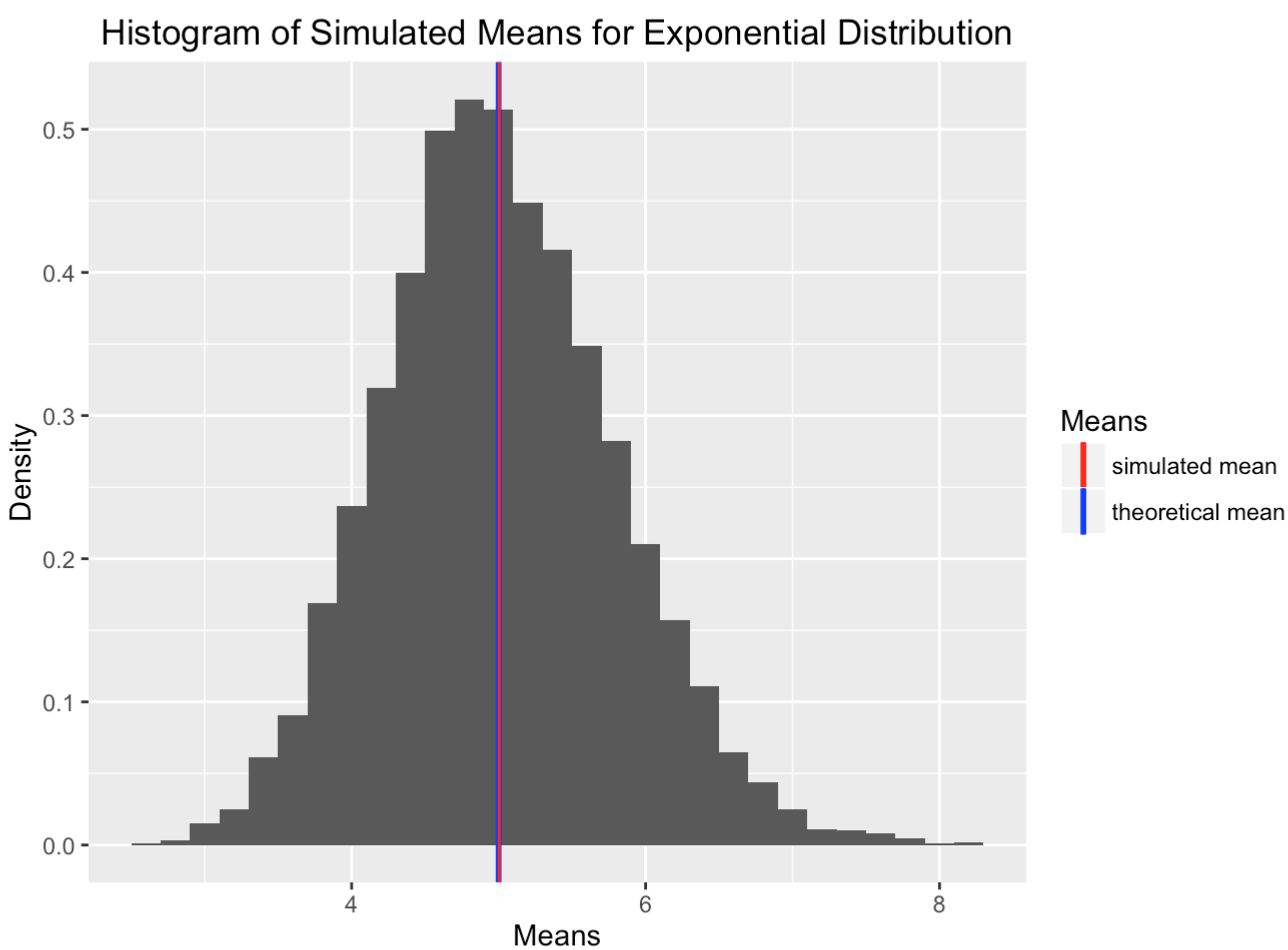
```
sim.mean <- mean(exp.dist.sim.means)
```

The value of `sim.mean` is: 5.0043033, which is extremely close to our theoretical mean of 5!

Let's plot a histogram with both the theoretical mean and the simulation mean.

```
# create a dataframe of the means for use in ggplot
df.sim.means <- data.frame(means = exp.dist.sim.means)

cols <- c("red", "blue")
names(cols) <- c("Theoretical mean", "Simulation mean")
p.means <- ggplot(df.sim.means, aes(means))
p.means + geom_histogram(binwidth = lambda, aes(y = ..density..)) +
        geom_vline(aes(xintercept = 1/0.2, color = "theoretical mean"), size = 1) +
        geom_vline(aes(xintercept = mean(df.sim.means$mean), color = "simulated mean"))
+
        scale_color_manual(name = "Means", values = c("red", "blue")) +
        labs(x = "Means", y = "Density", title = "Histogram of Simulated Means for Expon
ential Distribution")
```



Sample Variance vs Theoretical Variance

Now let's take a look at the theoretical variance versus the variance we see in our simulation. The theoretical standard deviation is `1/lambda/sqrt(n) = 1/0.2/sqrt(5000)` = 0.7905694 using our values.

The theoretical variance is: `1/lambda/sqrt(n)^2` = 0.625.

Let's compare these theoretical values to our simulation:

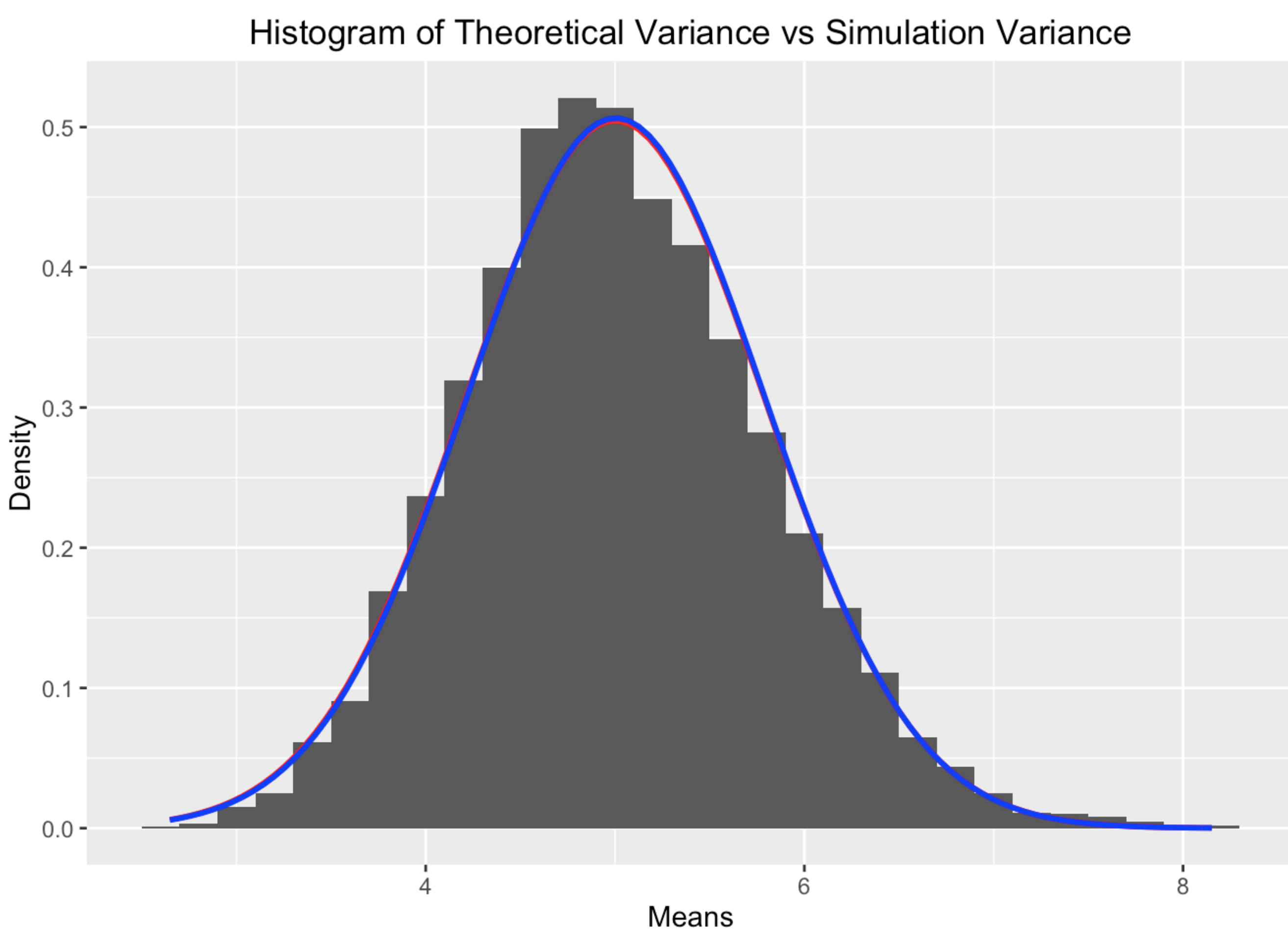
```
sd.means <- sd(exp.dist.sim.means)
var.sim.means <- sd.means ^ 2
```

In our simulation, we get a standard deviation of: 0.7877829, which is extremely close to our theoretical standard deviation of 0.7905694. And we get a variance of 0.620602, which is very close to our theoretical value of 0.625.

Let's visualize using a plot. First we will put the theoretical statistics into variables:

```
sd.theoretical <- 1/lambda/sqrt(n)
mean.theoretical <- 1/lambda
```

```
p.var <- ggplot(df.sim.means, aes(means))
p.var + geom_histogram(binwidth = lambda, aes(y = ..density..)) +
        stat_function(fun=dnorm,args=list(mean=mean.theoretical, sd=sd.theoretical), col
or = "red", size = 1.0) +
        stat_function(fun=dnorm,args=list(mean=sim.mean, sd=sd.means), color = "blue", s
ize = 1.0) +
        labs(x = "Means", y = "Density", title = "Histogram of Theoretical Variance vs S
imulation Variance")
```

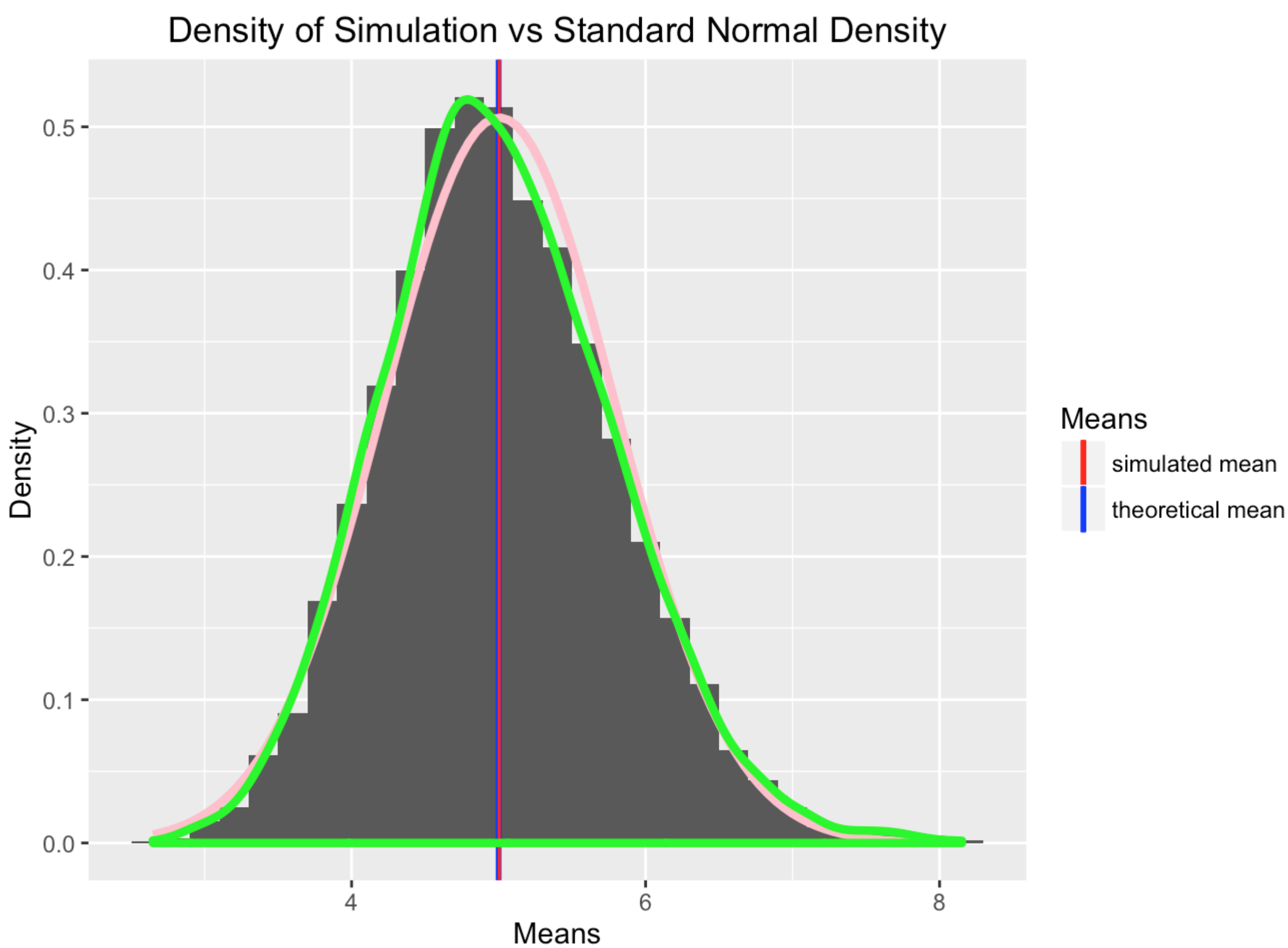


The red line corresponds to our theoretical variance whereas the blue line corresponds to the simulation variance. As we can see, the lines are nearly on top of each other, showing that the simulation data does, in fact, align well with the theory as we expected.

Distribution

In order to see if our simulated data is approximately normal, we will plot the simulated mean and standard deviation against a normal density distribution.

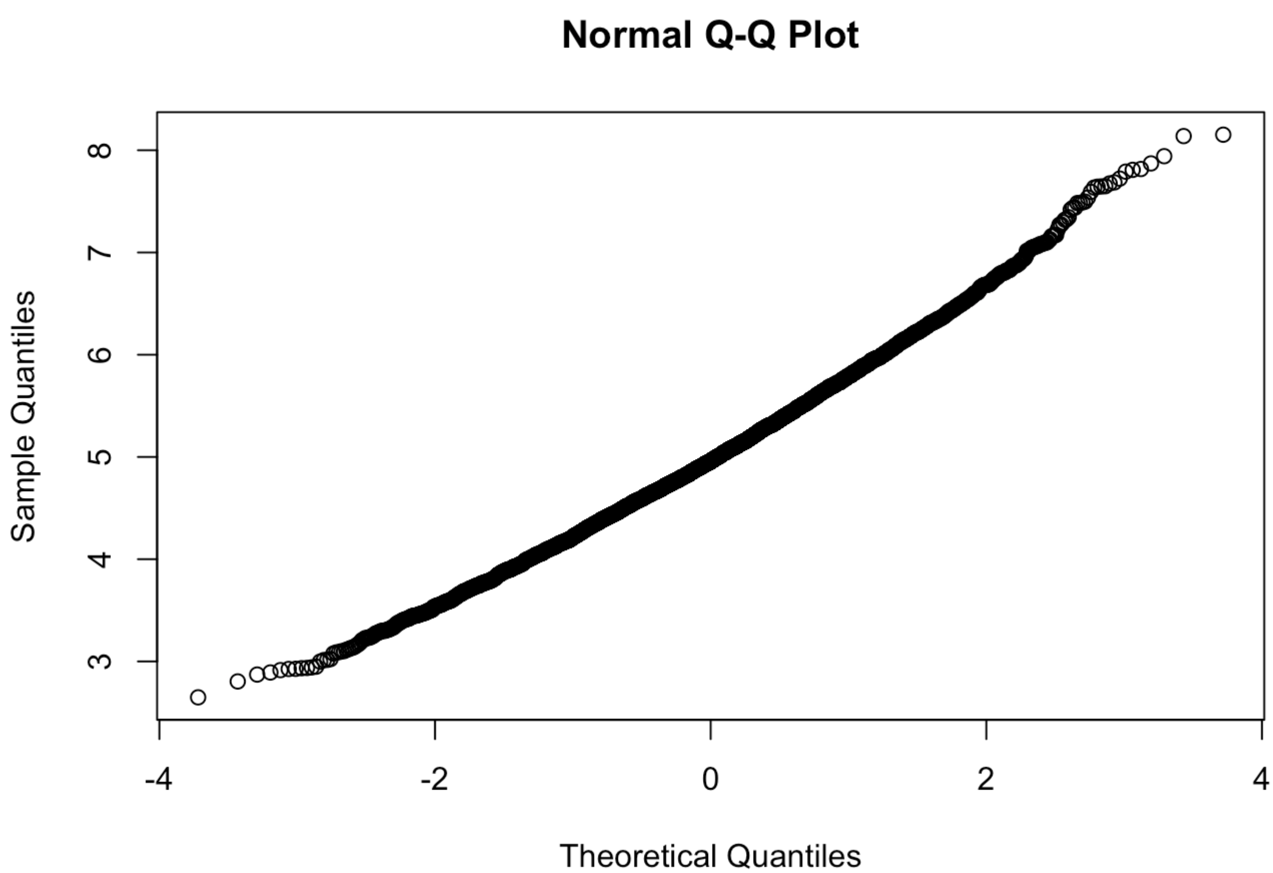
```
p.dist <- ggplot(df.sim.means, aes(means))
p.dist + geom_histogram(binwidth = lambda, aes(y = ..density..)) +
        geom_vline(aes(xintercept = 1/0.2, color = "theoretical mean"), size = 1) +
        geom_vline(aes(xintercept = mean(df.sim.means$mean), color = "simulated mean"))
+
        scale_color_manual(name = "Means", values = c("red", "blue")) +
        stat_function(fun=dnorm,args=list(mean=sim.mean, sd=sd.means), color = "pink", s
ize = 1.5) +
        geom_density(color = "green", size = 1.5) +
        labs(x = "Means", y = "Density", title = "Density of Simulation vs Standard Norm
al Density")
```



The green density line corresponds to a normal distribution while the pink line corresponds to our simulation data. As we can see the lines are very close to each other, which indicate strongly that our simulated data is approximately normal.

We can do a final check by using a Q-Q Plot with our simulated data:

```
qqnorm(df.sim.means$means)
```



The fact that our simulated data is almost perfectly stright shows that our simulated data is approximately normal.