Statistical Inference Project Part 1: Exponential Distribution v.s. Central Limit Theorem

Joselle Abagat

5/27/2018

Criteria

1. Did you show where the distribution is centered at and compare it to the theoretical center of the distribution?
2. Did you show how variable it is and compare it to the theoretical variance of the distribution?
3. Did you perform an exploratory data analysis of at least a single plot or table highlighting basic features of the data?
4. Did the student perform some relevant confidence intervals and/or tests?
5. Were the results of the tests and/or intervals interpreted in the context of the problem correctly?
6. Did the student describe the assumptions needed for their conclusions?

Load libraries

The R libraries used in this report are: ggplot2, datatable, DT

Part 1: Exponential Distribution v.s. Central Limit Theorem

In this project you will investigate the exponential distribution in R and compare it with the Central Limit Theorem.

According to wikipedia, “the exponential distribution is the probability distribution that describes the time between events in a Poisson point process, i.e., a process in which events occur continuously and independently at a constant average rate.”

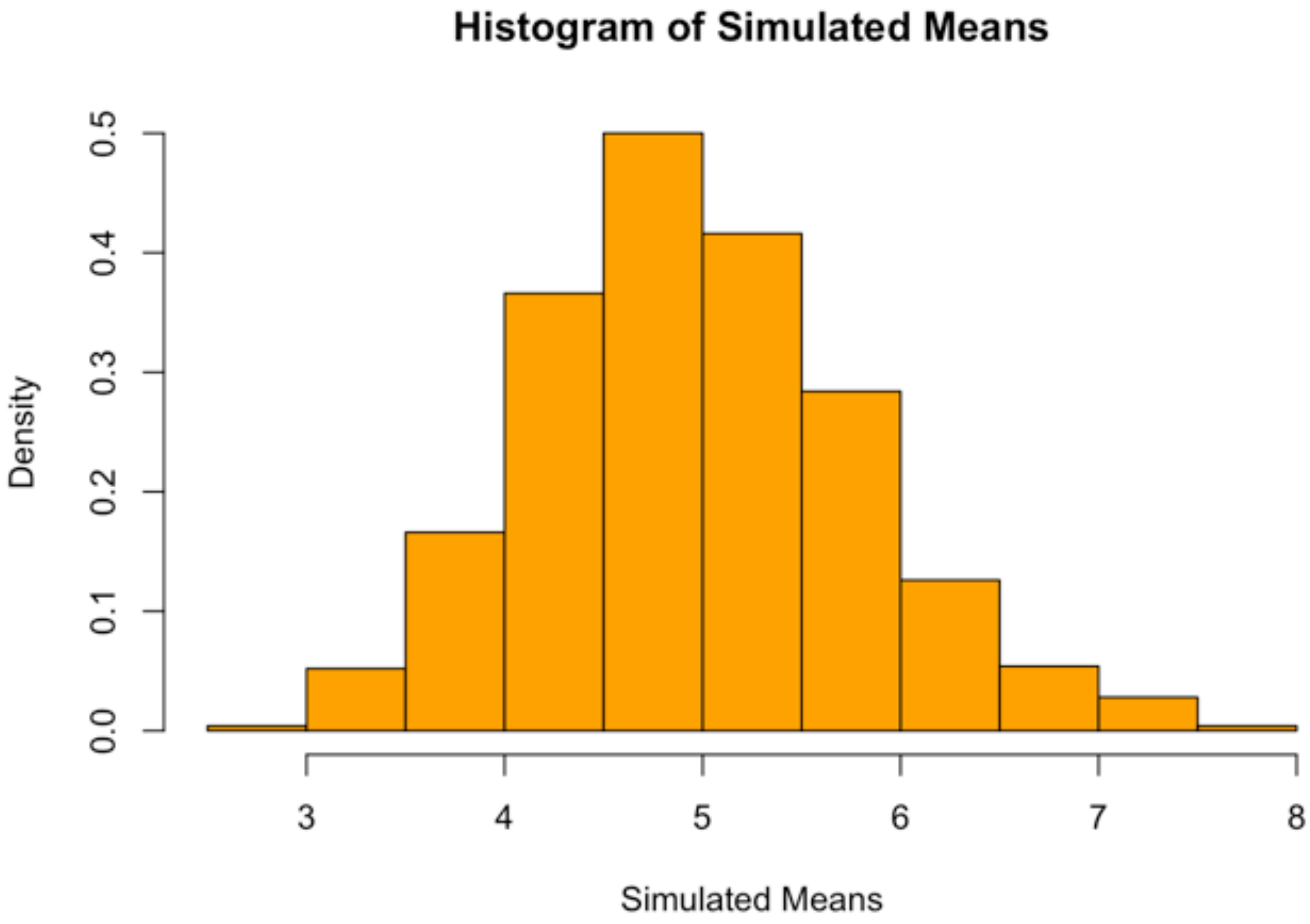
The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Set lambda = 0.2 for all of the simulations. You will investigate the distribution of averages of 40 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.

```
# Declare Variables
lambda <- 0.2 #rate parameter
sd <- 1/lambda
nExp <- 40
nSim <- 1000
# set.seed(123) #set

# Create data frame matrix
dt <- data.table(matrix(rexp(n = nSim*nExp, rate = lambda), nrow = nSim, ncol = nExp)
)
# Obtain means
# data = dt, margin = 1 (perform by row), use function mean
dtMeans <- data.table(apply(dt, 1, mean))

# Illustrate
hist(dtMeans$V1, col = "orange", freq = FALSE, main = "Histogram of Simulated Means",
xlab = "Simulated Means")
```



1. Show the sample mean and compare it to the theoretical mean of the distribution.

The mean of exponential distribution is 1/lambda:

```
1/lambda
```

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

theoretical mean = 1/lambda = 5. Looking at the histogram, we can see that the simulated mean is centered around 4.8-5.0. This can be further confirmed by looking at the summary of means where the median and mean are very close to the theoretical mean.

```
summary(dtMeans)
```

```
##        V1
##  Min.   :2.669
##  1st Qu.:4.397
##  Median :4.901
##  Mean   :4.968
##  3rd Qu.:5.492
##  Max.   :7.749
```

2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.

Sample Variance = s^2/n

```
# Sample Variance
sd^2/nExp
```

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

```
# Theoretical Variance
var(dtMeans$V1)
```

```
## [1] 0.6634484
```

As we can see, the sample and the theoretical variances are very close in value.

3. Show that the distribution is approximately normal.

The Central Limit Theorem (CLT) states that the distribution of averages of iid variables (properly normalized) becomes that of a standard normal as the sample size increases. The equations below show that the normalization of the sample and theoretical values are approximately equal.

```
meanSample <- 1/lambda
meanTheoretical <- mean(dtMeans$V1)
# Theoretical
(meanTheoretical - meanSample)/sqrt(var(dtMeans$V1))
```

```
## [1] -0.03963217
```

```
# Sample
(meanTheoretical - meanSample)/(sd/sqrt(nExp))
```

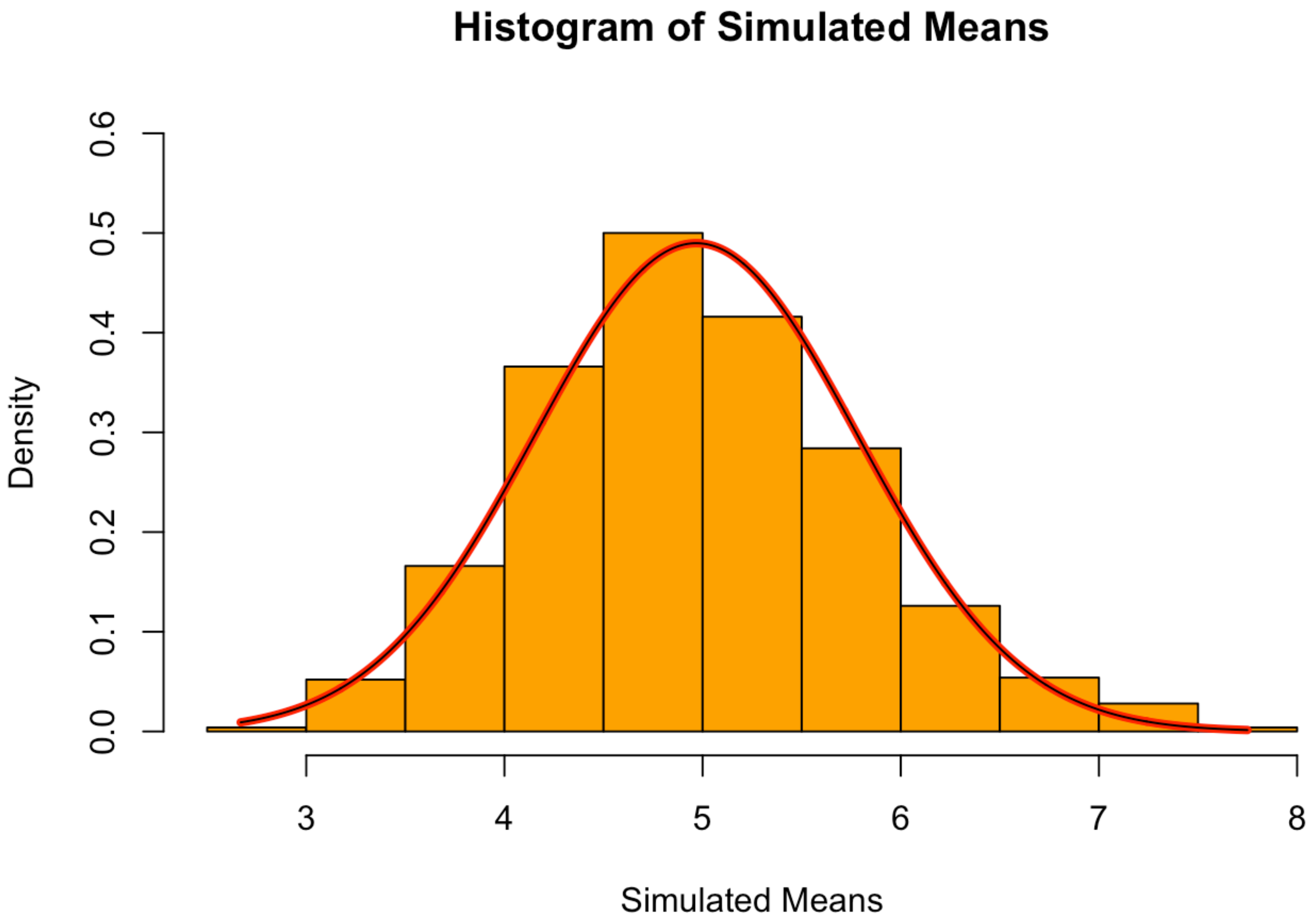
```
## [1] -0.04083301
```

Let’s increase the sample size and look at what happens to the simulation. By the CLT, as the sample size increases, we get closer to a standard normal distribution.

```
nMore <- 100
dtNMore <- data.table(matrix(rexp(n = nSim*nExp, rate = lambda), nrow = nSim, ncol =
nExp))
dtMeansNMore <- data.table(apply(dt, 1, mean))
```

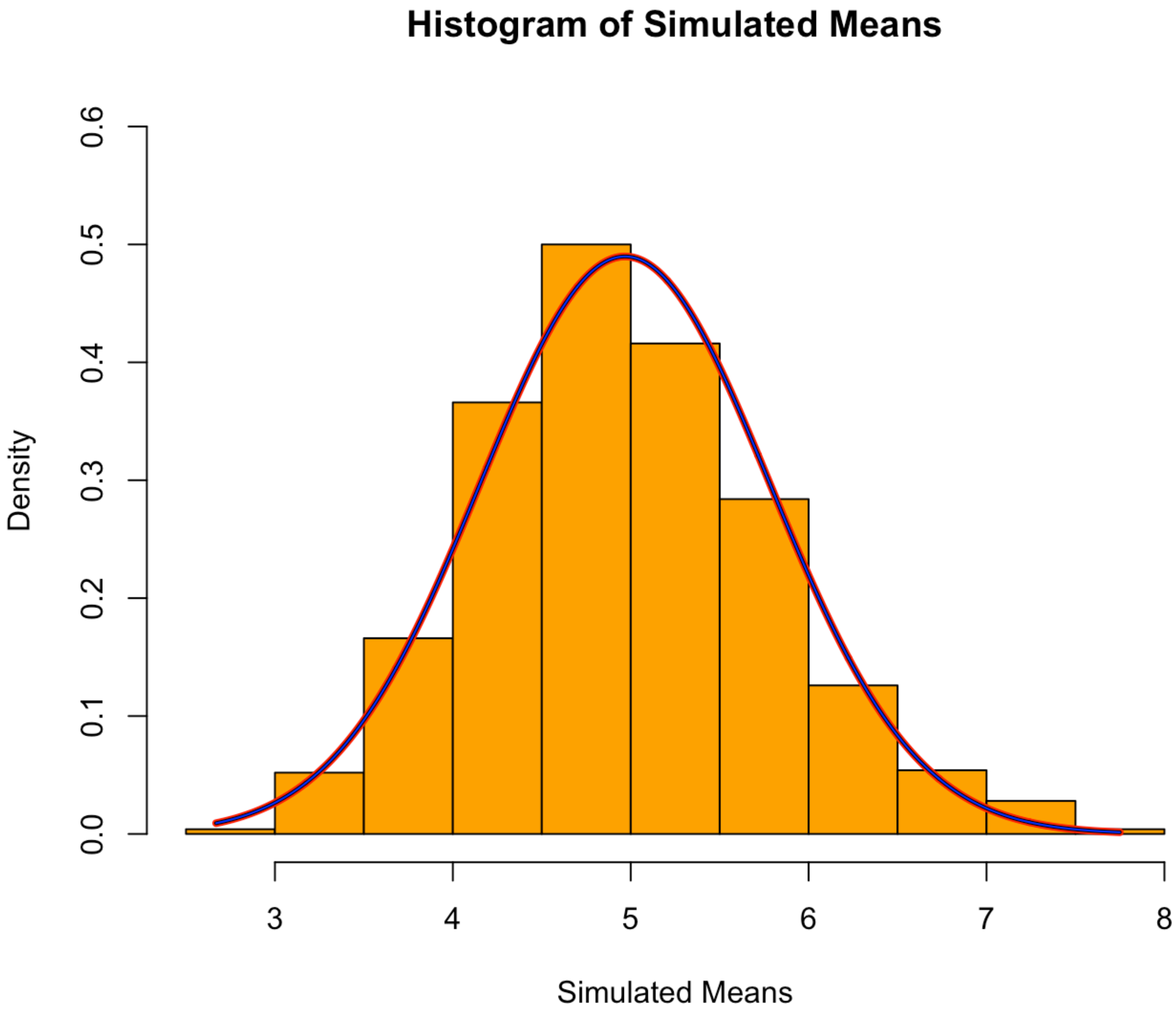
Now, let’s plot the increased sample size and compare it to the original sample size. The plot below shows that the distribution is approximately normal: the red line represents n = 40 and the black line represents n = 100. As we can see, both distributions are approximately normal.

```
hist(dtMeansNMore$V1, col = "orange", freq = FALSE, main = "Histogram of Simulated Me
ans", xlab = "Simulated Means", ylim = c(0, 0.6))
xfit <- seq(min(dtMeans$V1), max(dtMeans$V1), length = nrow(dtMeans))
yfit <- dnorm(xfit, mean = mean(dtMeans$V1), sd = sd(dtMeans$V1))
lines(xfit, yfit, col = "red", lwd = 4)
xfit2 <- seq(min(dtMeansNMore$V1), max(dtMeansNMore$V1), length = nrow(dtMeansNMore))
yfit2 <- dnorm(xfit, mean = mean(dtMeansNMore$V1), sd = sd(dtMeansNMore$V1))
lines(xfit2, yfit2, col = "black", lwd = 1)
```



Let’s try increasing the sample size even more and compare the distributions: the red line represents n = 40, the black line represents n = 100, and the blue line represents n = 1000. All distributions are approximately normal.

```
nEvenMore <- 1000
dtNEvenMore <- data.table(matrix(rexp(n = nSim*nExp, rate = lambda), nrow = nSim, ncol = nExp))
dtMeansNEvenMore <- data.table(apply(dt, 1, mean))
hist(dtMeansNEvenMore$V1, col = "orange", freq = FALSE, main = "Histogram of Simulated Means", xlab = "Simulated Means", ylim = c(0, 0.6))
xfit <- seq(min(dtMeans$V1), max(dtMeans$V1), length = nrow(dtMeans))
yfit <- dnorm(xfit, mean = mean(dtMeans$V1), sd = sd(dtMeans$V1))
lines(xfit, yfit, col = "red", lwd = 4)
xfit2 <- seq(min(dtMeansNMore$V1), max(dtMeansNMore$V1), length = nrow(dtMeansNMore))
yfit2 <- dnorm(xfit, mean = mean(dtMeansNMore$V1), sd = sd(dtMeansNMore$V1))
lines(xfit2, yfit2, col = "black", lwd = 2)
xfit3 <- seq(min(dtMeansNEvenMore$V1), max(dtMeansNEvenMore$V1), length = nrow(dtMeansNEvenMore))
yfit3 <- dnorm(xfit, mean = mean(dtMeansNEvenMore$V1), sd = sd(dtMeansNEvenMore$V1))
lines(xfit2, yfit2, col = "blue", lwd = 1)
```



Per the CLT, the simulation is approximately normal as we increased the sample size.