SI Course Project

Yigit Ozan Berk

6/27/2019

Overview

Part 1 : 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.

Part 2: Now in the second portion of the project, we're going to analyze the ToothGrowth data in the R datasets package.

Load the ToothGrowth data and perform some basic exploratory data analyses Provide a basic summary of the data. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there's other approaches worth considering) State your conclusions and the assumptions needed for your conclusions.

Part 1

Simulations

Initiation of required packages:

```
library(ggplot2)
```

Simulating 1000 sets of 40 exponential random variables:

```
set.seed(121212)
my_lambda <- 0.2
my_n <- 40
nsim <- 1000

samples <- matrix(nrow = 40, ncol = 1000)
for(i in 1:1000) samples[,i] = rexp(my_n, my_lambda)
```

Means, Standard Deviations, and Variances of simulated 40 exponential random variables:

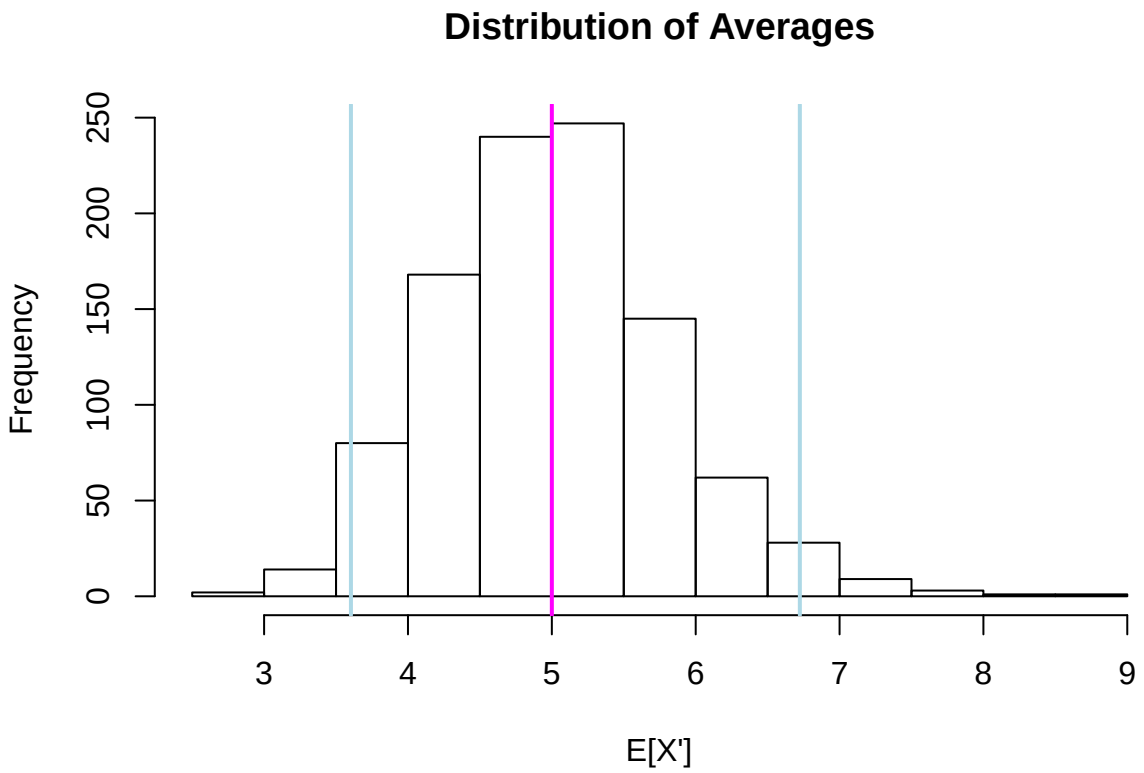
```
mns <- apply(samples, 2, mean)

sdevs <- apply(samples, 2, sd)

vars <- apply(samples, 2, var)
```

Sample Mean versus Theoretical Mean

The theoretical mean of the distribution is lambda = 1/0.2 = `1/.2`

The overall average of 1000 simulated 40 exponential random variables are `mean(mns)`

```
hist(mns, main = "Distribution of Averages",
     xlab = "E[X']")
abline(v = 5, col = "magenta", lwd = 2)
abline(v = quantile(mns, c(.025, .975)), col = "lightblue", lwd = 2)
```



The distribution of averages is centered around the theoretical mean as expected; with 95% confidence interval:

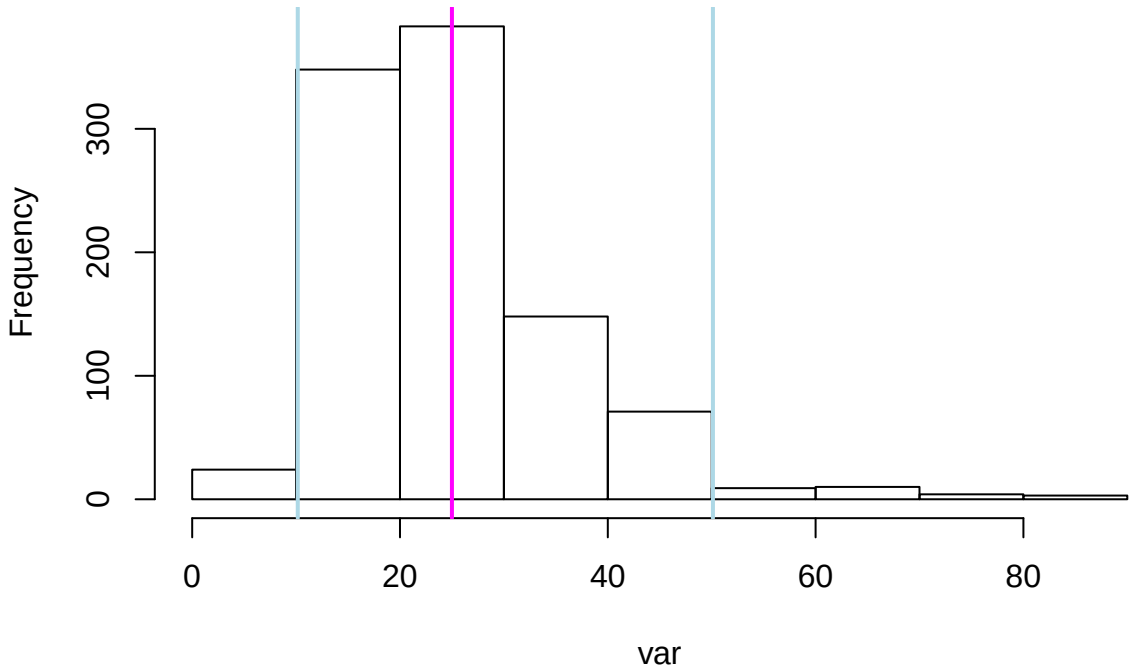
```
quantile(mns, c(.025, .975))
```

```
##     2.5%    97.5%
## 3.603528 6.723835
```

Sample Variance versus Theoretical Variance

With a theoretical variance of 25 *(1/lambda^2)*, the distribution of observed variances are as follows

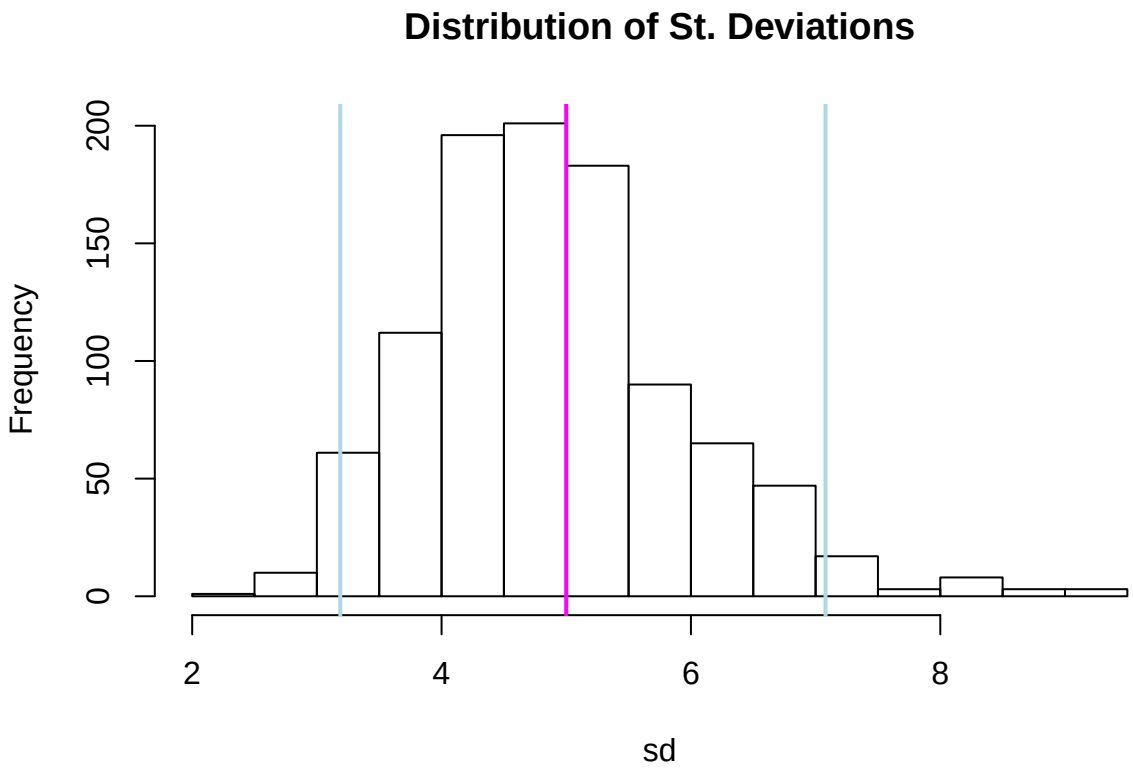
```
hist(vars, main = "Distribution of Variances",
     xlab = "var")
abline(v = 25, col = "magenta", lwd = 2)
abline(v = quantile(vars, c(.025,.975)), col = "lightblue", lwd = 2)
```



The average variation in the simulations is observed as 25.03814 which is close to theoretical variance.

If we look at the distribution of standard deviations, we can see that it's centered at the theoretical standard deviation with an approximately normal distribution:

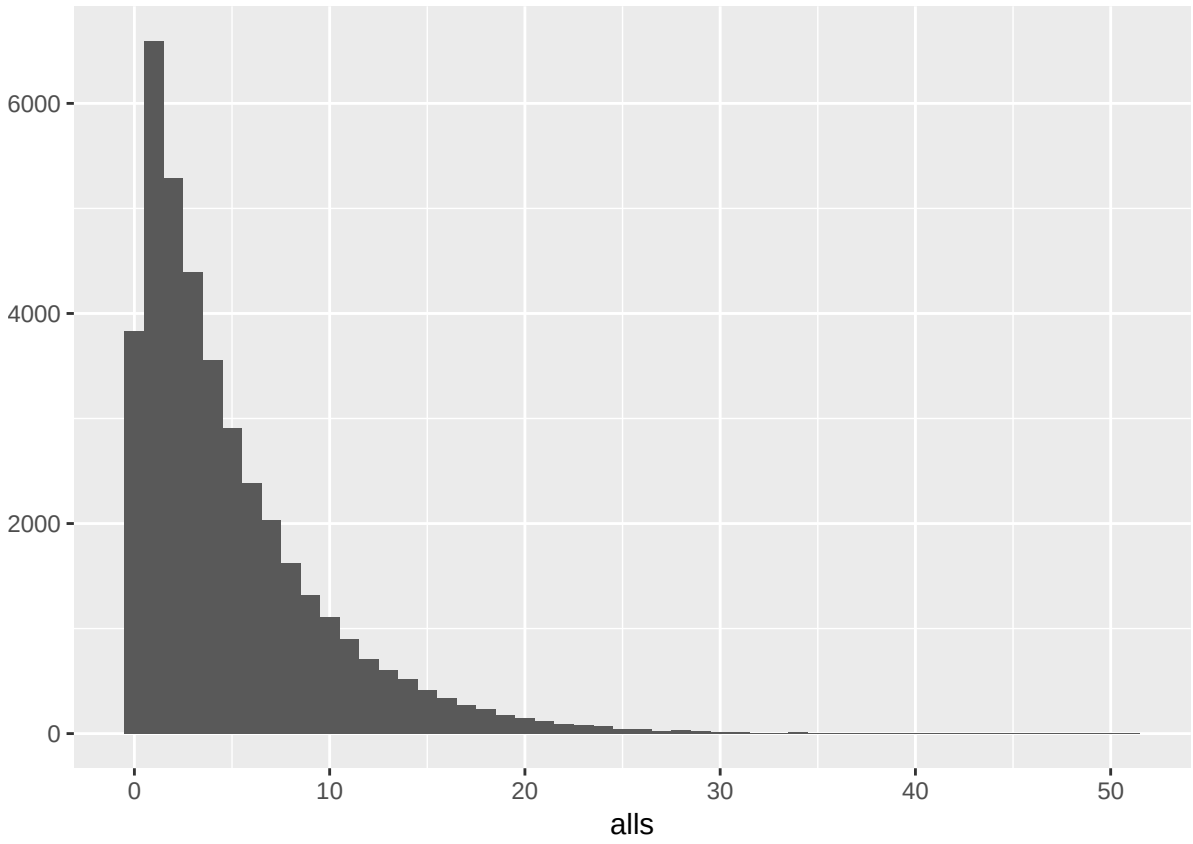
```
hist(sdevs, main = "Distribution of St. Deviations",
     xlab = "sd")
abline(v = 5, col = "magenta", lwd = 2)
abline(v = quantile(sdevs, c(.025,.975)), col = "lightblue", lwd = 2)
```



Distribution

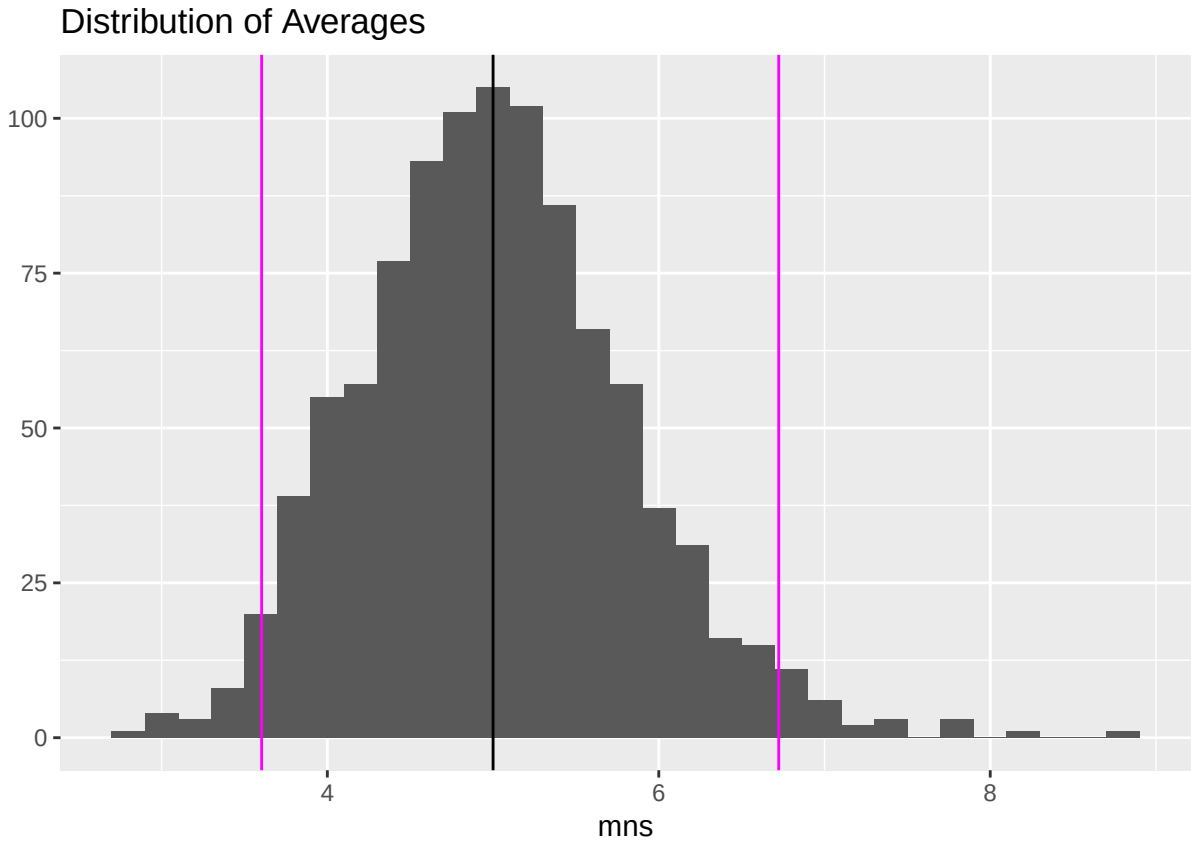
If we look at all the random draws of the exponential random variables, we can see the graph of the frequencies constitutes an exponential distribution. By CLT as the size of the variables increases, the frequency density will transform into a finer exponential distribution density graph.

```
alls <- NULL
for(i in 1:40) {
        alls <- c(alls, samples[i, ])
}
qplot(alls, binwidth = 1)
```



However, if we look at the distribution of averages of 40 exponential random variables over the 1000 simulations, we can clearly see the frequency distribution is not equal to an exponential distribution:

```
g <- qplot(mns, binwidth = .2)
g <- g + geom_vline(xintercept = 5)
g <- g + geom_vline(xintercept = quantile(mns, c(.025, .975)), col = "magenta")
g <- g + ggtitle("Distribution of Averages")
g
```



Instead, the distribution of averages is roughly a normal distribution with mean centered around the theoretical mean(5) of the population. One can tell the distribution is roughly normal by the bell shape curve of the frequency density.

By CLT as the number of simulations increases, the distribution of averages of these 40 random draws will look more and more like a normal distribution with mean equal to theoretical mean of the population.