Statistical Inference Course Project 1

Ziyao Gao

10/30/2017

Assignment Description

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:

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.

Enviroment requirement for this report:

```
# install the pacakes needed
library(knitr)
library(ggplot2)
# set seed to make the report reproducible
set.seed(12345)
```

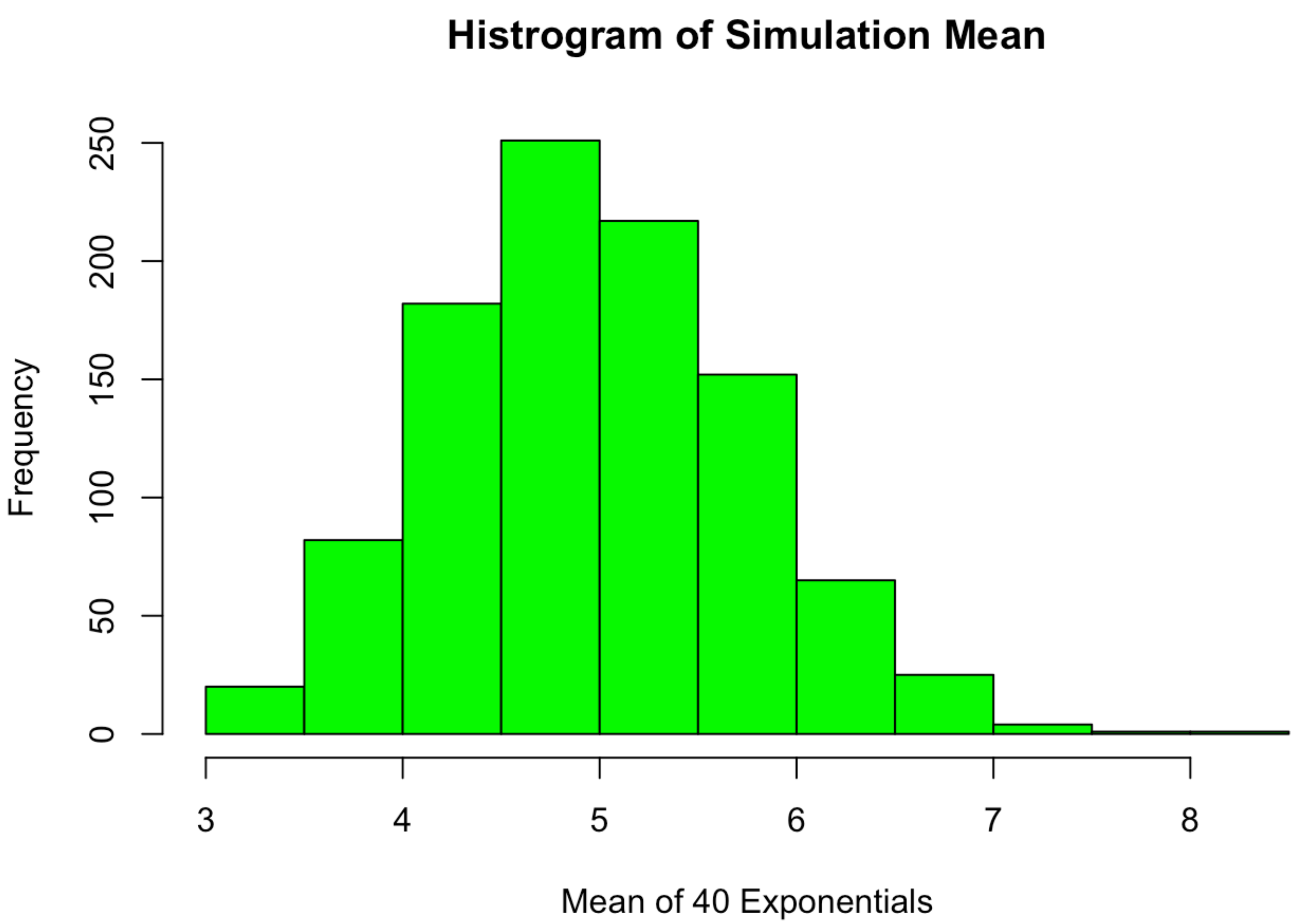
Simulation Exercise

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.

```
# set variables for the simulation
n <- 40
lambda <- 0.2
simulations <- 1000

# Create a matrix of 1000 rows with the columns for random simulation 40 times
matrix <- matrix(rexp(simulations * n, rate = lambda), simulations, n)

# calcualte means and plot it through histogram
sim_mean <- rowMeans(matrix)
hist(sim_mean, xlab = "Mean of 40 Exponentials", ylab = "Frequency",
     main = "Histrogram of Simulation Mean", col = "green")
```



Sample Mean vs. Theoretical Mean

```
sample_mean <- mean(sim_mean)
sample_mean
```

```
## [1] 4.971972
```

```
theoretical_mean <- 1/lambda
theoretical_mean
```

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

As we can see above, sample mean is really close to theoretical mean.

Sample Variance vs. Theoretical Variance

```
sample_var <- var(sim_mean)
sample_var
```

```
## [1] 0.6157926
```

```
theoretical_var <- (1/lambda)^2/n
theoretical_var
```

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

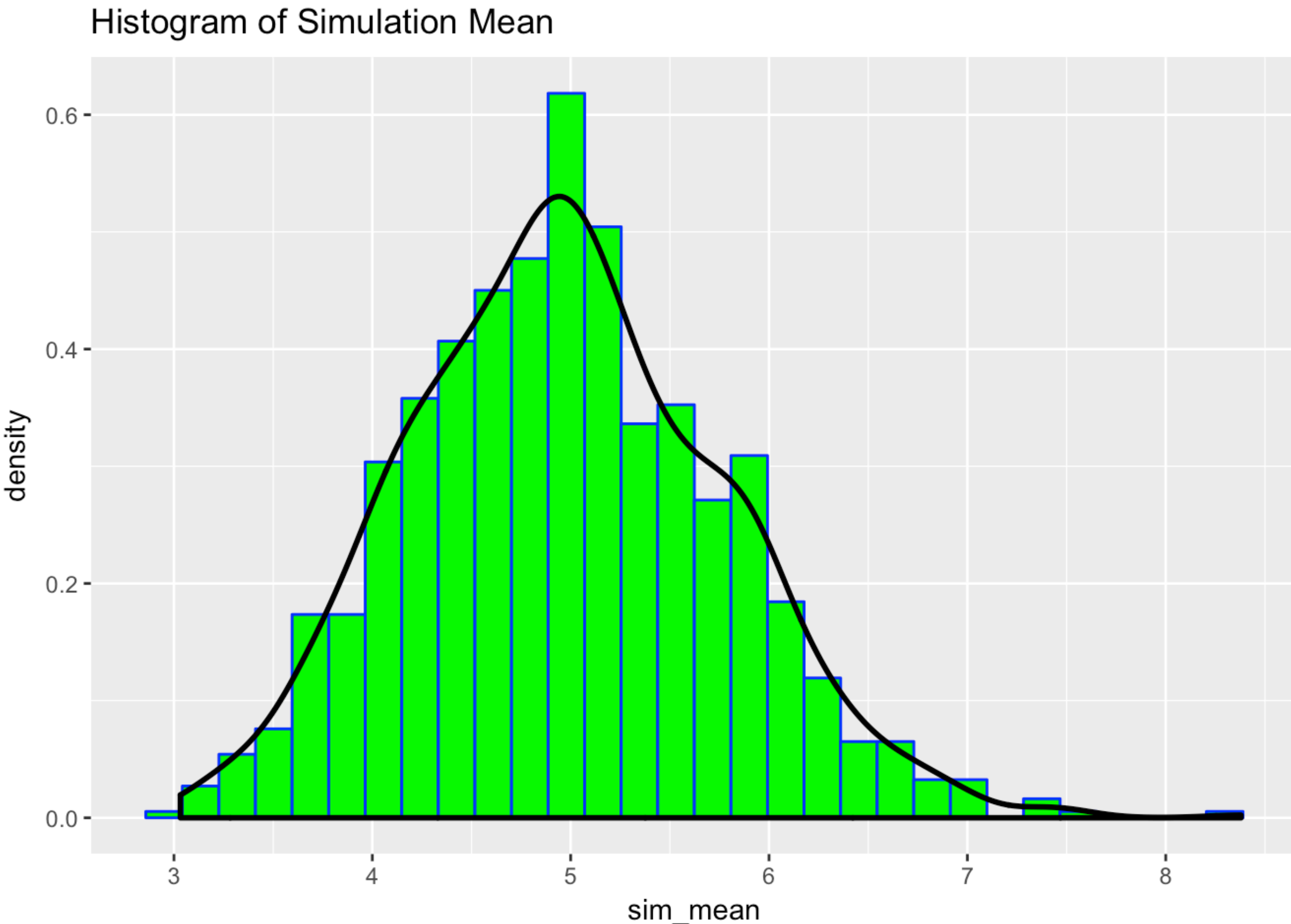
As we can see above, sample variance is also very close to theoretical variance.

Approximate normal distribution

First, create an approximate normal distribution and see how the sample aligns with it

```
plotdata <- data.frame(sim_mean)
g <- ggplot(plotdata, aes(x = sim_mean))
g = g + geom_histogram(aes(y = ..density..), colour = "blue", fill = "green")
g = g + geom_density(colour = "black", size = 1)
g = g + ggtitle("Histogram of Simulation Mean")
g
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The plot above indicates that the histogram can be adeqautely approximated with the normal distribution.

Second, let's compare their confidence intervals as well:

```
sample_conf_interval <- round(mean(sim_mean) + c(-1,1)*1.96*sd(sim_mean)/sqrt(n), 3)
sample_conf_interval
```

```
## [1] 4.729 5.215
```

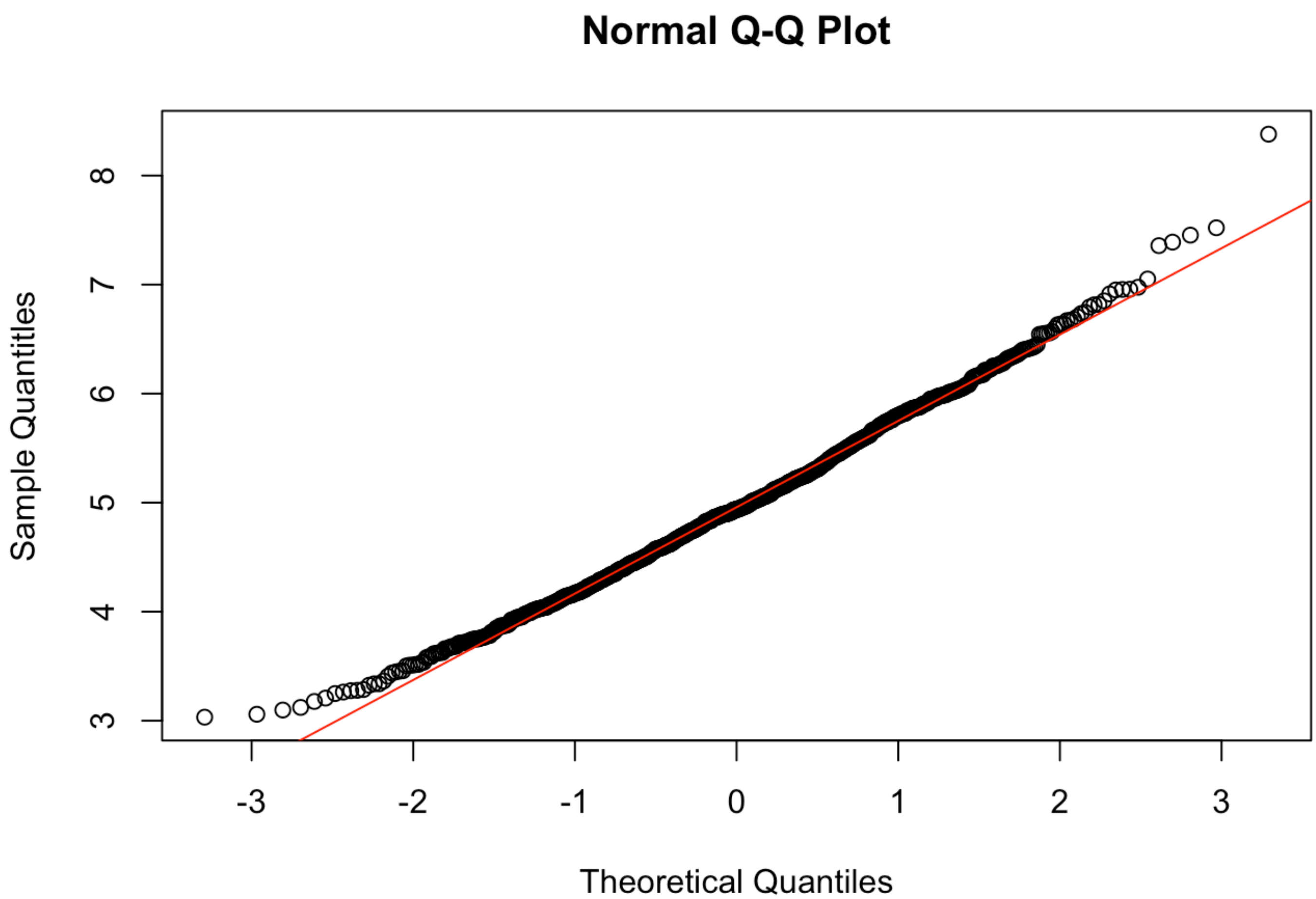
```
theoretical_conf_interval <- round(mean(theoretical_mean) + c(-1, 1)*1.96*sqrt(theore
tical_var)/sqrt(n), 3)
theoretical_conf_interval
```

```
## [1] 4.755 5.245
```

From the results we can see, the sample 95% confidence interval [4.729, 5.215] is close to theoretical confidence interval [4.755, 5.245] as well

Third, plot Q-Q for quantitles.

```
qqnorm(sim_mean, main = "Normal Q-Q Plot", xlab = "Theoretical Quantiles", ylab = "Sa
mple Quantitles")
qqline(sim_mean, col = "red")
```



As we can see, the distribuion is approximately normal.