Project Report for Simulation Exercise

Yukai Zou

Overview

This report is the first part of the Coursera statistical inference course project. In this report, an exponential distribution is simulated, and the distribution of averages of 40 exponentials is investigated. Explained through figures and explanatory texts, the simulated sample mean is compared with the theoretical mean, the simulated sample variance is compared with the theoretical variance, and whether the distribution is approximately normal is discussed.

Sample Mean versus Theoretical Mean

Sample mean:

```
## [1] 4.982267
```

Theoretical mean:

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

The sample mean is pretty close to the theoretical mean. As is shown in Figure 1 (See Appendix for reference), the distribution of 1000 simulations looks like the Gaussian distribution (whose mean is 5.0 and variance is 0.625), with the sample mean located near the theoretical mean of (highlighted in blue color).

Sample Variance versus Theoretical Variance

Sample variance:

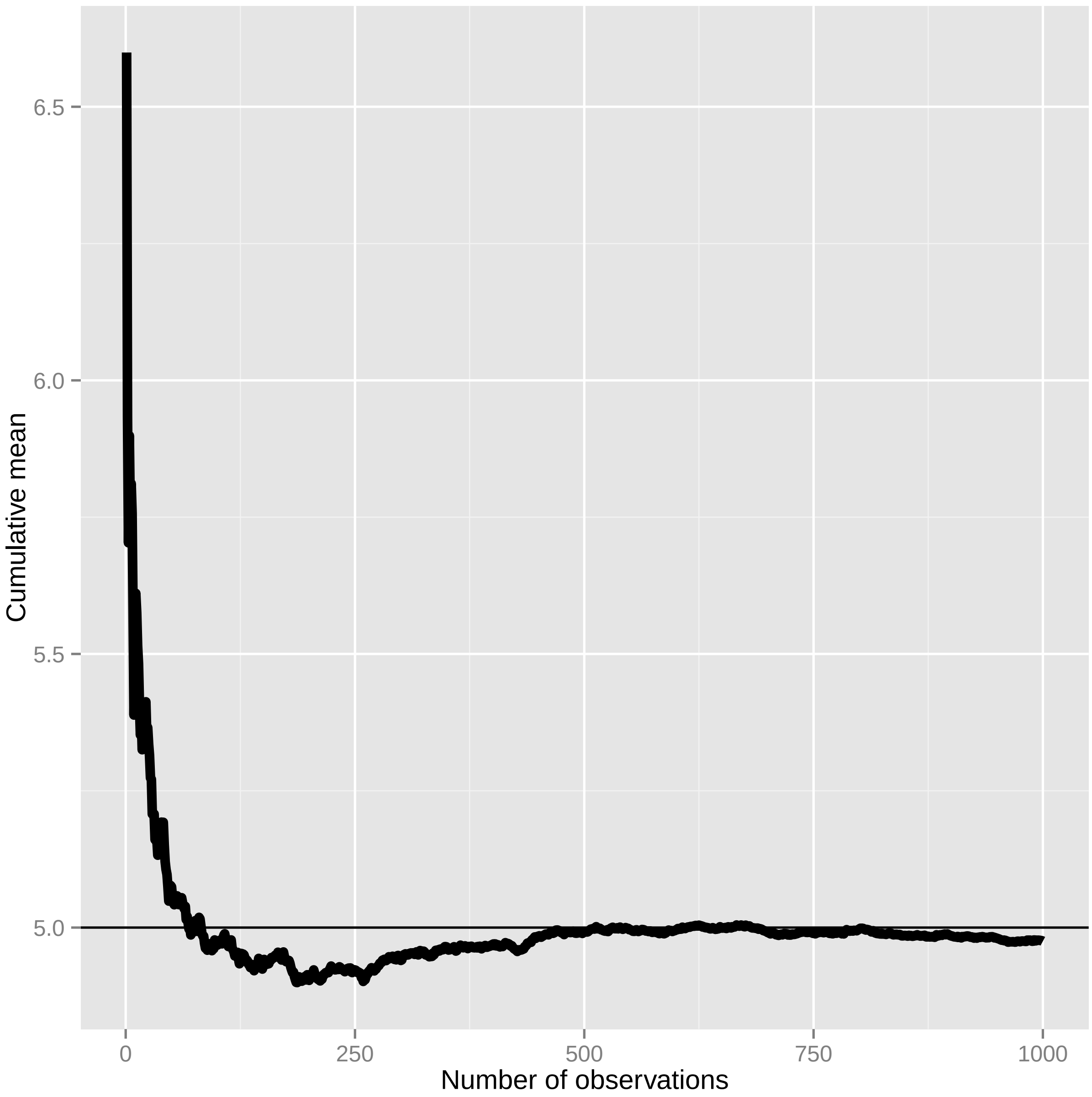
```
## [1] 0.6535137
```

Theoretical variance:

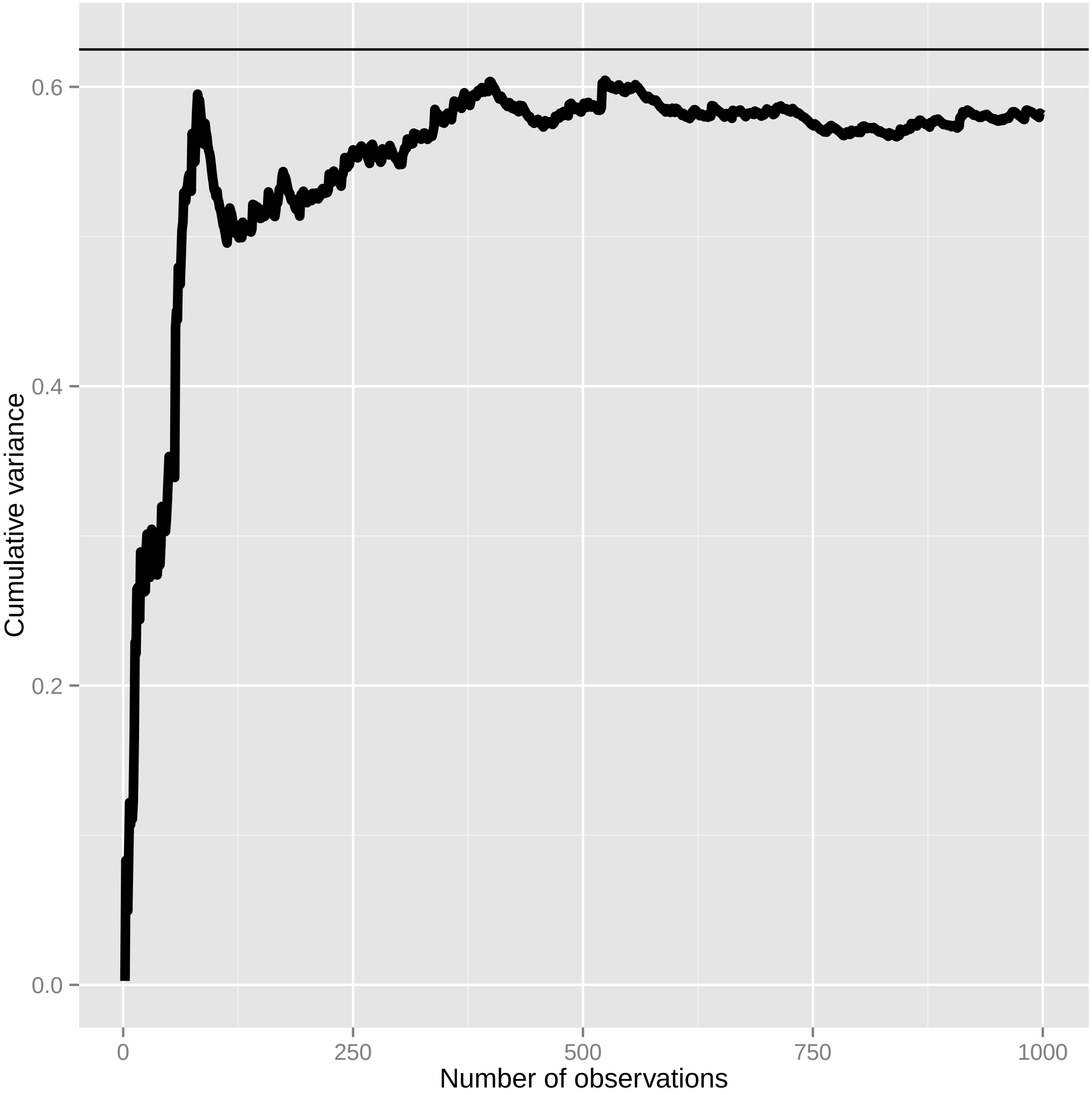
```
## [1] 0.625
```

The sample variance is quite close to the theoretical variance. As is shown in Figure 2 (See Appendix for reference), there is not much difference between the standard deviation of the simulation and the theoretical standard deviation (highlighted in red lines) of the normal distribution curve, which indicates a good estimation of theoretical variance using the sample variance from 1000 simulations.

Discussion: Is the Distribution Approximately Normal?



As the number of simulation becomes larger, the estimated sample mean is getting closer to the value of theoretical mean (5.0, highlighted by the horizontal line) with a little fluctuation, which is consistent with the Law of large numbers.



As the number of simulation becomes larger, the estimated sample variance has some fluctuations but is getting more and more closer to the value of theoretical variance (0.625, highlighted by the horizontal line), which is consistent with the Law of large numbers.

According to Central Limit Theorem, when the sample size increases in a simulation, the distribution of iid (independent and identically distributed) variables will become more and more approximately normal, in which the distribution will be approaching the population mean, and the standard deviation will be approaching the standard error of the mean. In our simulation example, the average of 40 exponentials is assumed to be an iid variable, and 1000 averages can be considered as a large collection of data. Therefore, the distribution of 1000 averages of 40 exponentials is approximately normal.

Appendix

R codes

```
# Initialization and Simulation
library(ggplot2) # Load the ggplot2 package
mns <- NULL # Initialize a NULL variable mns
for (i in 1:1000) { # Run a for loop for 1000 times
        mns <- c(mns, mean(rexp(40, rate = 0.2)))
        # Run the random generation for the exponential
        # distribution with 40 observations and a rate
        # of 0.2. After that, the average of the 40
        # observations is calculated, concatenated with
        # the mns variable to update the previous mns.
        }

# Plot Figure 1 in Appendix Section
hist(mns, xlim = c(2,8), main = "Results of one thousand simulations",
     xlab = "x", ylab = "Count")
abline(v = mean(mns), lwd = 3)
x <- seq(2, 8, length = 1000)
y <- dnorm(x, mean = 5, sd = 5/sqrt(40))
par(new = T)
plot(x, y, type = "l", lwd = 3, xaxt = "n", yaxt = "n",
     xlab = "", ylab = "", col = "blue")
axis(4, col = "blue", col.axis = "blue")
abline(v = mean(mns), col = "blue", lwd = 3)
legend("topright", pch = 20, col = c("black", "blue"),
       c("Simulations", "Normal distribution"), cex = 0.7)

# Plot Figure 2 in Appendix Section
hist(mns, xlim = c(2,8))
x <- seq(2, 8, length = 1000)
y <- dnorm(x, mean = 5, sd = 5/sqrt(40))
par(new = T)
plot(x, y, type = "l", xaxt = "n", yaxt = "n",
     xlab = "", ylab = "", col = "red")
axis(4, col = "red", col.axis = "red")
ssd <- 5 + c(-1, 1) * sqrt(var(mns))
abline(v = ssd)
tsd <- 5 + c(-1, 1) * sqrt(5^2/40)
abline(v = tsd, col = "red")
legend("topright", pch = 20, col = c("black", "red"),
       c("Simulations", "Normal distribution"), cex = 0.65)

# Plot cumulative mean of the simulations in Distribution Section
meanmns <- NULL
varmns <- NULL
for (i in 1:1000) {
        mns <- c(mns, mean(rexp(40, rate = 0.2)))
        meanmns <- c(meanmns, mean(mns))
        varmns <- c(varmns, var(mns))
}
g1 <- ggplot(data.frame(x = 1:1000, y = meanmns), aes(x=x, y=y))
```

```
g1 <- g1 + geom_hline(yintercept = 5.0) + geom_line(size = 2)
g1 <- g1 + labs(x = "Number of observations", y = "Cumulative mean")
g1

# Plot cumulative variance of the simulations in Distribution Section
g2 <- ggplot(data.frame(x = 2:1000,
             y = varmns[2:1000]), aes(x=x, y=y)) # Omit the first y value since it is NA
g2 <- g2 + geom_hline(yintercept = 5^2/40) + geom_line(size = 2)
g2 <- g2 + labs(x = "Number of observations", y = "Cumulative variance")
g2
```

Simulation Figures

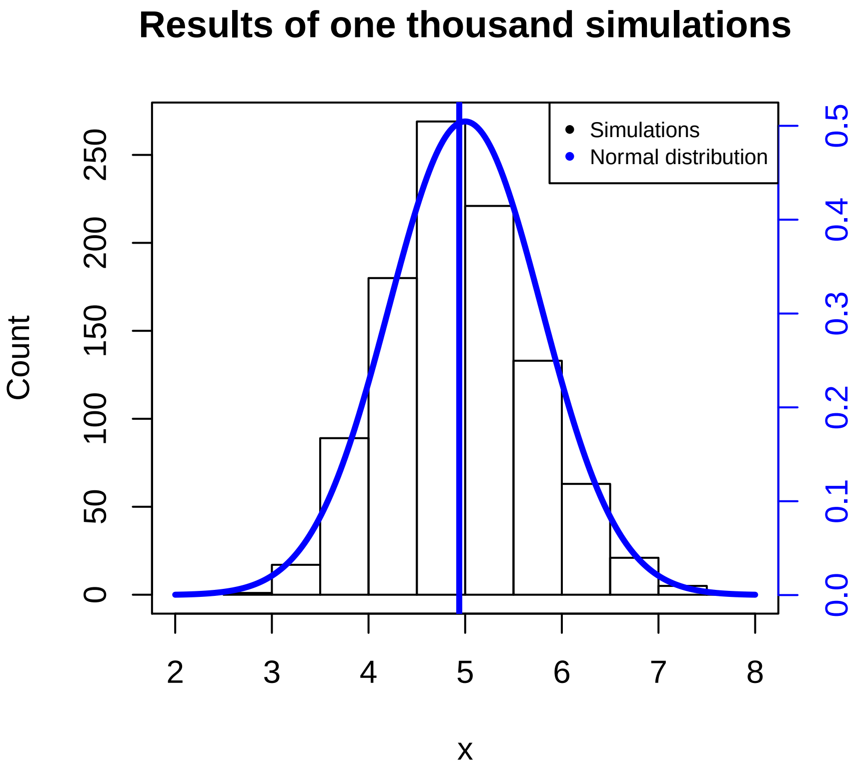


Figure 1: Comparisons of sample mean and theoretical mean

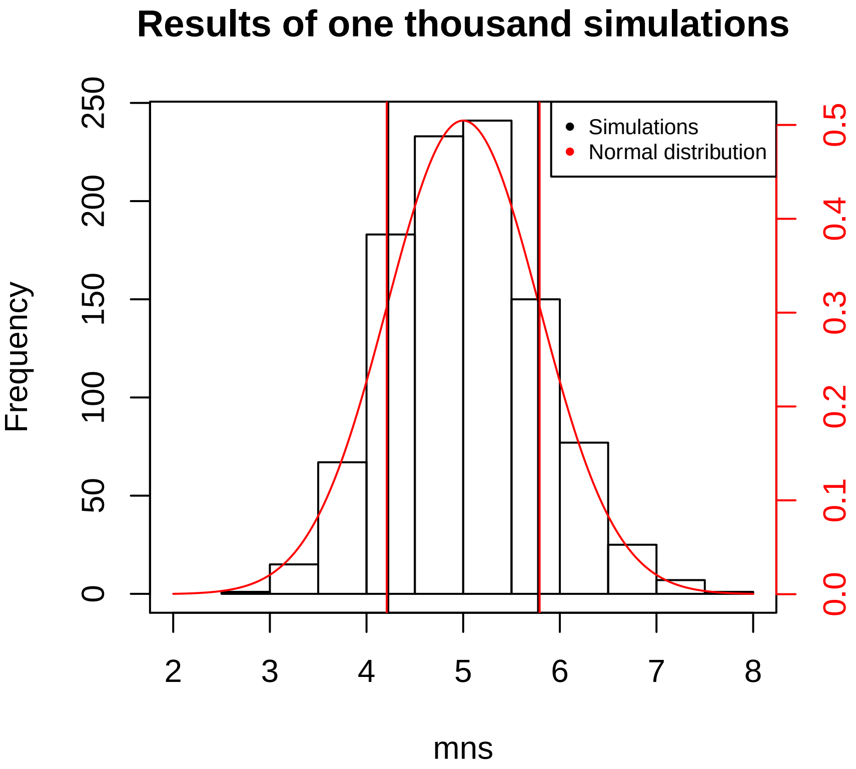


Figure 2: Comparisons of sample variance and theoretical variance