Project Report for Simulation Exercise

Yukai Zou

Overview

This report is the first part of the project for the statistical inference class. In this report, an exponential distribution will be simulated, and the distribution of averages of 40 exponentials will be investigated. Using figures and explanatory texts, the simulated sample mean will be compared with theoretical mean, the simulated sample variance will be compared with theoretical variance, and whether the distribution is approximately normal will also be discussed.

Sample Mean versus Theoretical Mean

Sample mean:

```
## [1] 4.986336
```

Theoretical mean:

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

As is shown in Figure 1 (See Appendix for reference), the distribution of 1000 simulations looks like a Gaussian distribution, and the sample mean is very close to the theoretical mean of the normal distribution curve (highlighted in blue color).

Sample Variance versus Theoretical Variance

Sample variance:

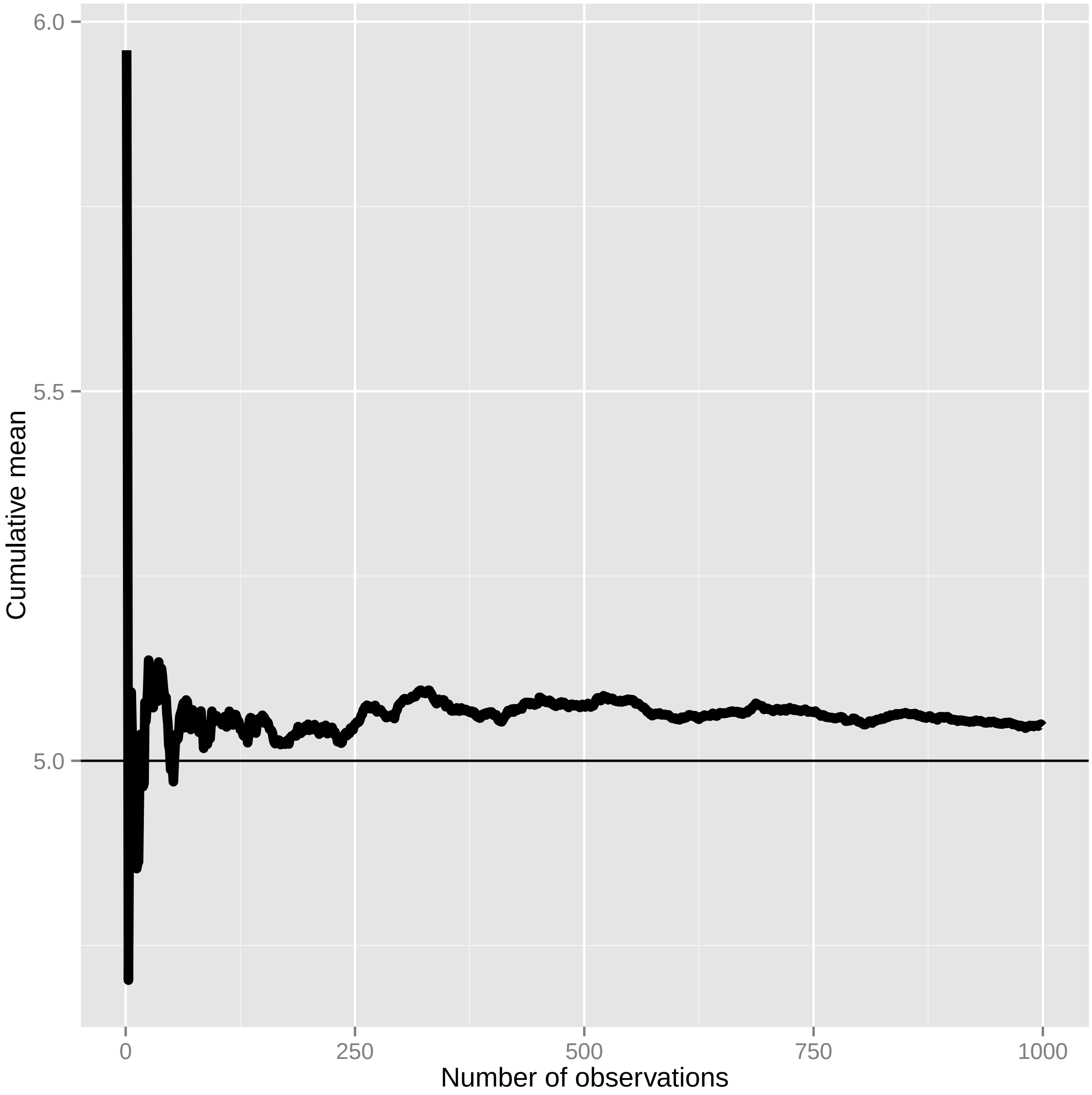
```
## [1] 0.6413181
```

Theoretical variance:

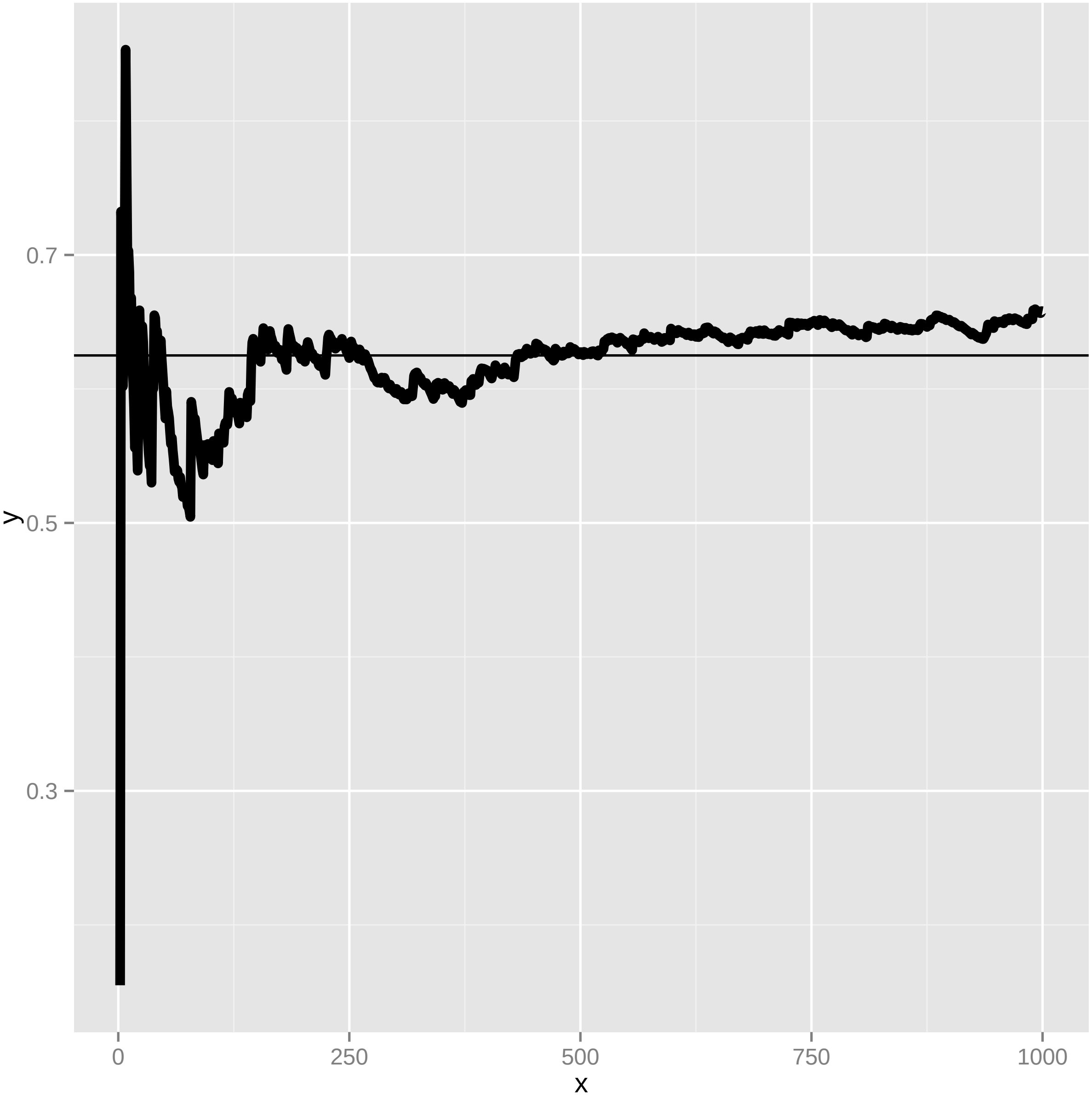
```
## [1] 0.625
```

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 of the normal distribution curve (highlighted in red color), indicating a good estimation of theoretical variance using the sample variance from 1000 simulations.

Distribution



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



As the number of simulation becomes larger, the estimated sample variance is getting more and more closer to the value of theoretical variance (0.625), which is consistent with the Law of large numbers.

The averages of 40 exponentials from the 1000 random exponentials are assumed to be iid (independent and identically distributed). According to Central Limit Theorem, for the simulations, as the sample size increases, the distribution of iid variables will become standard normal, and the distribution will be centered at the population mean with the standard deviation equal to the standard error of the mean.

Therefore, the distribution of 1000 averages of 40 exponentials is approximately normal.

Appendix

R codes

Include English explanations of the simulations you ran, with the accompanying R code. Your explanations should make clear what the R code accomplishes.

```r
library(ggplot2)
mns <- NULL
for (i in 1:1000) mns <- c(mns, mean(rexp(40, rate = 0.2)))
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)

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)

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

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
```

Figures

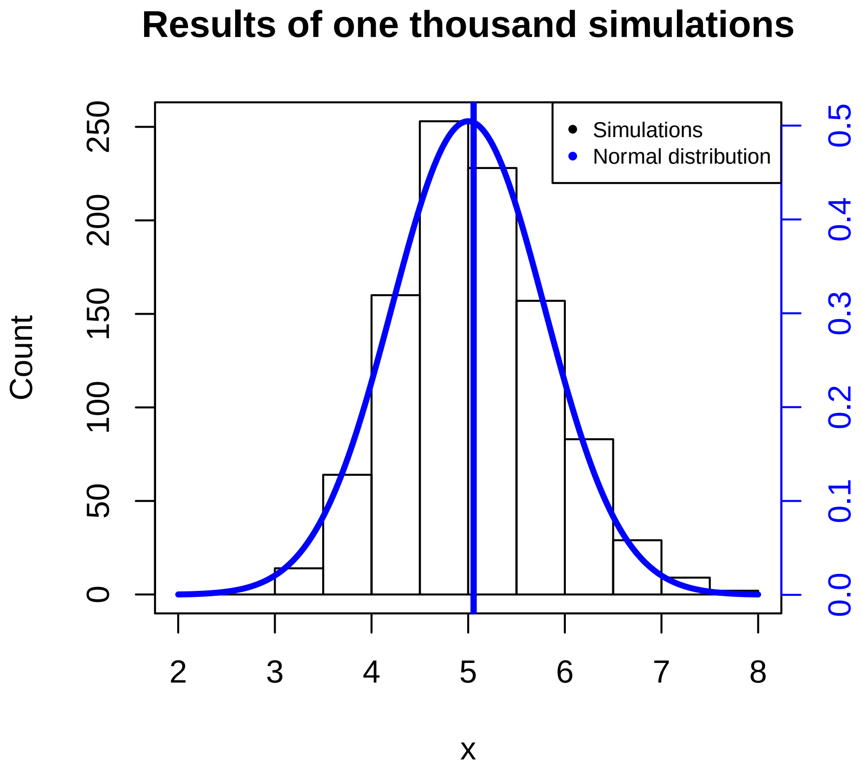


Figure 1: Comparisons of sample mean and theoretical mean

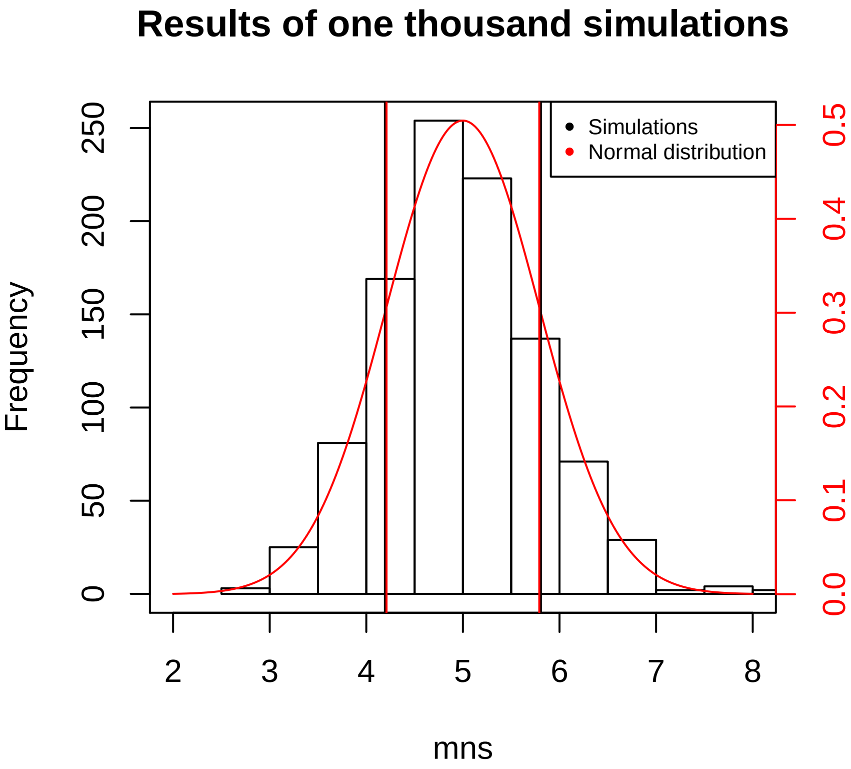


Figure 2: Comparisons of sample variance and theoretical variance