Inferential Data Analysis for Exponential Distribution

Statistical Inference Course project part 1

Ray Qiu

Overview

- In this project, we will use simulations to investigate the exponential distribution in R and compare it with the Central Limit Theorem.
- The report includes the follows:
  1. The sample mean and how it compares to the theoretical mean of the distribution.
  2. How variable the sample is (via variance) and how it compares to the theoretical variance of the distribution.
  3. The proof of that the distribution is approximately normal.

Load required R libraries

```
library(ggplot2)
```

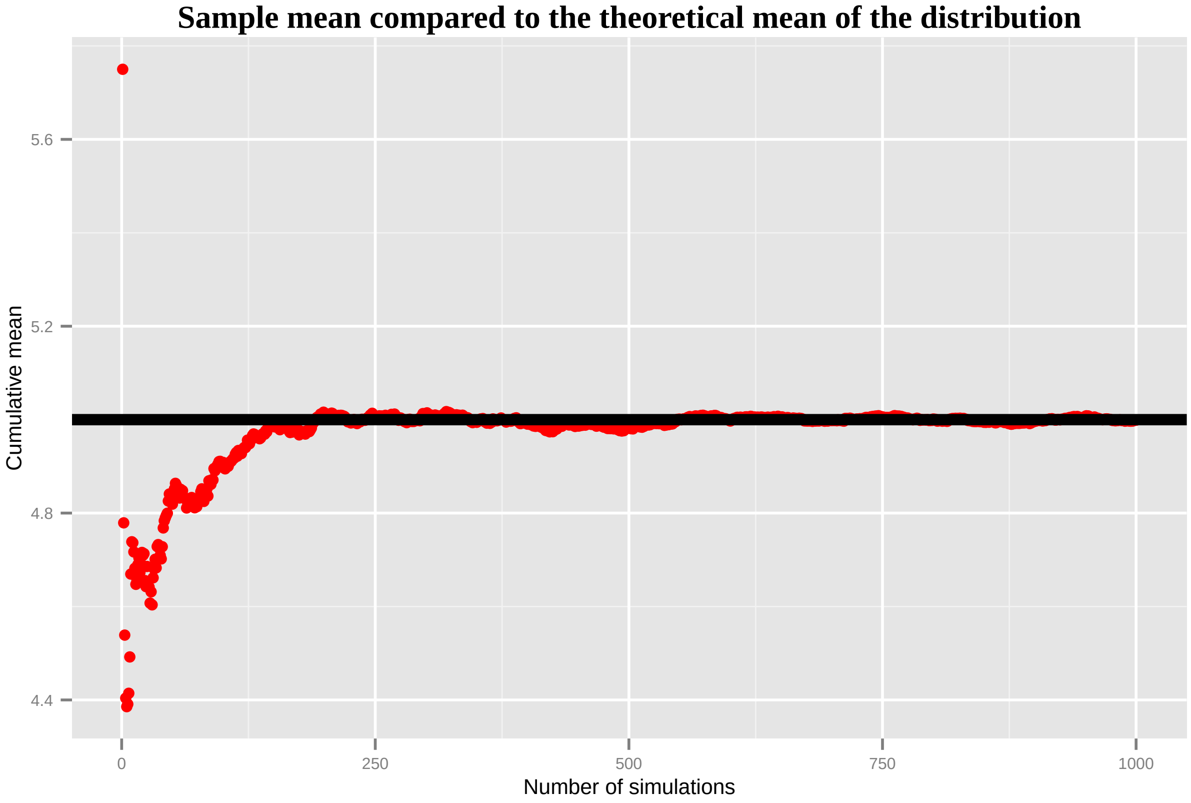
Simulations

- The exponential distribution is 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. We will investigate the distribution of averages of 40 exponentials in a total of 1000 simulations. The following R code generates the required data, and computes the means and save them to a dataframe.

```
lambda <- 0.2
n <- 1000
m <- 40
set.seed(820) # Set a seed for random number generation
dat <- replicate(n, mean(rexp(m, lambda)))
```

Problem #1: Compare the sample mean to the theoretical mean of the distribution.

```
means <- cumsum(dat)/(1:n)
ggplot(data.frame(x = 1:n, y = means), aes(x = x, y = y)) +
    geom_point(color = "red") +
    geom_hline(yintercept = 1 / lambda, size = 2) +
    labs(x = "Number of simulations", y = "Cumulative mean") +
    ggtitle("Sample mean compared to the theoretical mean of the distribution") +
    theme(text=element_text(size=8),
          plot.title=element_text(family="Times", face="bold", size=12))
```

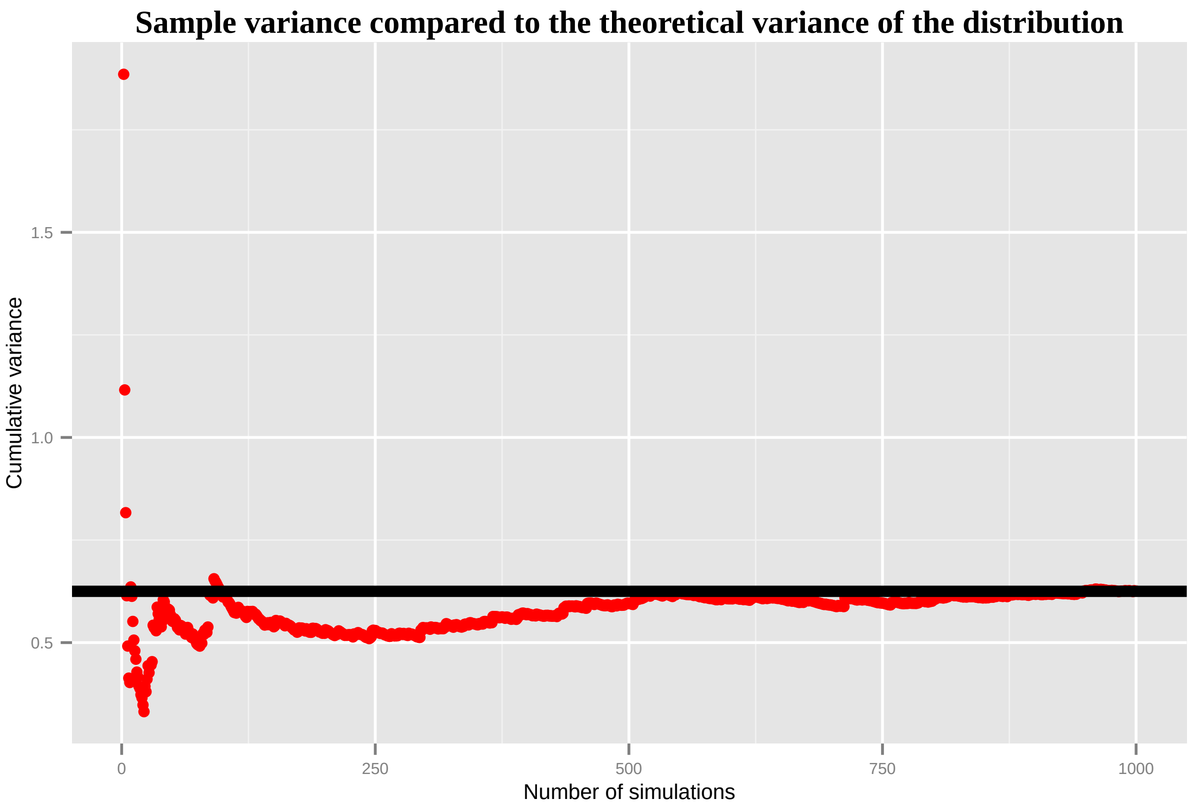


Solution #1: From the plot, we can see that as number of simulations goes up, sample mean estimates the theoretical mean of the distribution.

Problem #2: How variable the sample is (via variance) and how it compares to the theoretical variance of the distribution.

```
vars <- sapply(1:n, function(x) {
    var(head(dat, x))
})
ggplot(data.frame(x = 1:n, y = vars), aes(x = x, y = y)) +
    geom_point(color = "red") +
    geom_hline(yintercept = (1 / lambda) ^ 2 / m, size = 2) +
    labs(x = "Number of simulations", y = "Cumulative variance") +
    ggtitle("Sample variance compared to the theoretical variance of the distribution") +
    theme(text=element_text(size=8),
          plot.title=element_text(family="Times", face="bold", size=12))
```

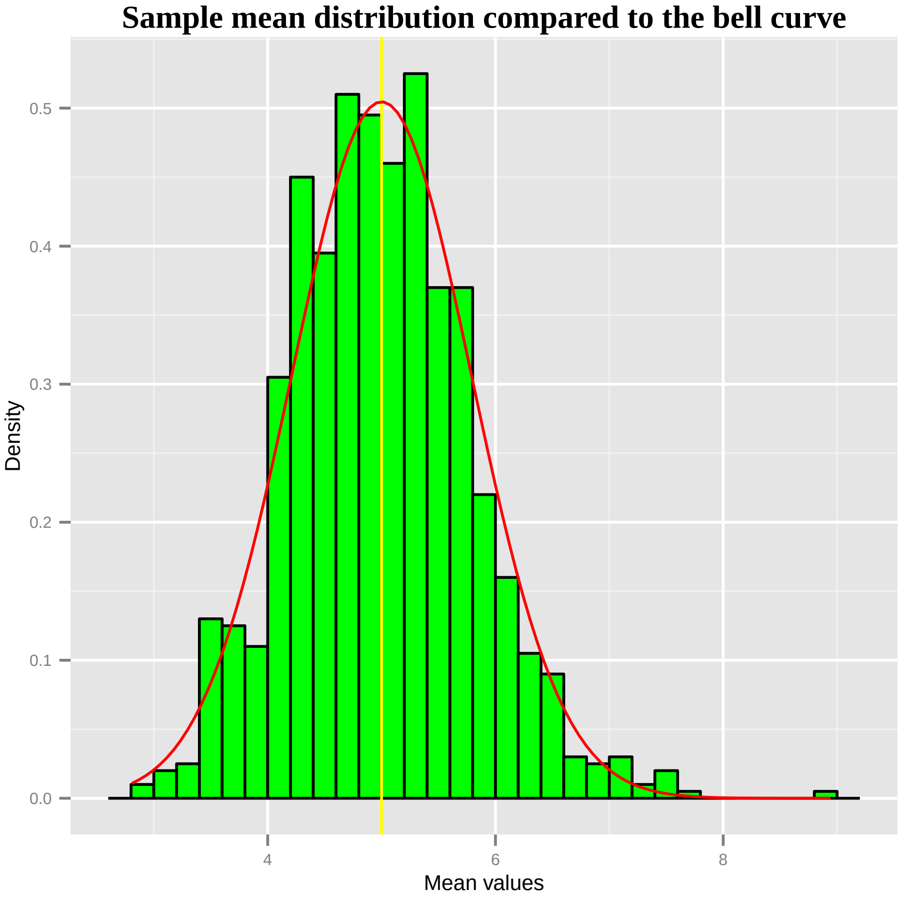
```
## Warning: Removed 1 rows containing missing values (geom_point).
```



Solution #2: From the plot, we can see that as number of simulations goes up, sample variance estimates the theoretical variance of the population.

Problem #3: Can we prove that the distribution of the sample mean is approximately normal?

```
ggplot(data.frame(x = dat), aes(x = x)) +
    geom_histogram(fill = "green", color = "black", binwidth = 0.2, aes(y = ..density..)) +
    labs(x = "Mean values", y = "Density") +
    geom_vline(xintercept = 5, color = "yellow") +
    stat_function(fun = dnorm, colour = "red", arg = list(mean = 5,
                                                          sd = (1/lambda)/sqrt(40))) +
    ggtitle("Sample mean distribution compared to the bell curve") +
    theme(text=element_text(size=8),
          plot.title=element_text(family="Times", face="bold", size=12))
```



Solution #3: We can see that the distribution of the sample mean is approximately normal, because it matches the dnorm bell curve very well.