Statistical Inference - The exponential distribution

Sabank

April 25, 2015

Synopsis: The basic goal of this assignment is to investigate the exponential distribution in R and compare it with the Central Limit Theorem. From the simulation exercise, we found that the sample mean, sample variance as well, of iid samples are respectively consistent for the population mean and variance. We also found that the distribution of averages of iid variables becomes that of a standard normal for the sample size of 40 exponentials.

1. INTRODUCTION

This is the project for the statistical inference class. In it, we use simulation to explore inference and do some simple inferential data analysis.

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 also 1/lambda. We set lambda = 0.2 for all of the simulations.

We will illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials by answering the following:

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

2. SIMULATION EXERCISE

```r
# load R library
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.1.3

# define constants
nosim <- 1000 # number of simulation
n <- 40 # size of the sample of popultation
lambda <- .2 # fixed rate parameter for exponential distribution
set.seed(2) # seed to allow reproducibility
```

## 2.1 Show Sample mean and compare it to theoretical mean of the distribution

```
# draw 'nosim' random sample of size 'n' from the exponential
distribution
data <- rexp(nosim * n, lambda)
# approximation of sample mean 'asm'
data_mtx <- matrix(data, nrow = nosim, ncol = n)
data_mns <- apply(data_mtx, 1, mean)
asm <- mean(data_mns)
# theoretical sample mean 'sm' for exponential distribution = 1 /
lambda
sm <- 1 / lambda
```

- approximation of sample mean = 5.0163562

- theoretical sample mean = 5

2.2 Show how variable the sample is and compare it to theoretical variance of the distribution

Variability of cumulated means:

```
# visualization by computing cumulated means of 1000 random
exponentials
df1 <- data.frame(obs = 1:nosim, asm = cumsum(rexp(nosim, lambda)) /
(1:nosim))
# plot
g1 <- ggplot(df1, aes(x = obs, y = asm))
g1 + geom_line() +
    geom_hline(yintercept = sm, color = "Red") +
    labs(x = "simulation", y = "cumulative mean")
```
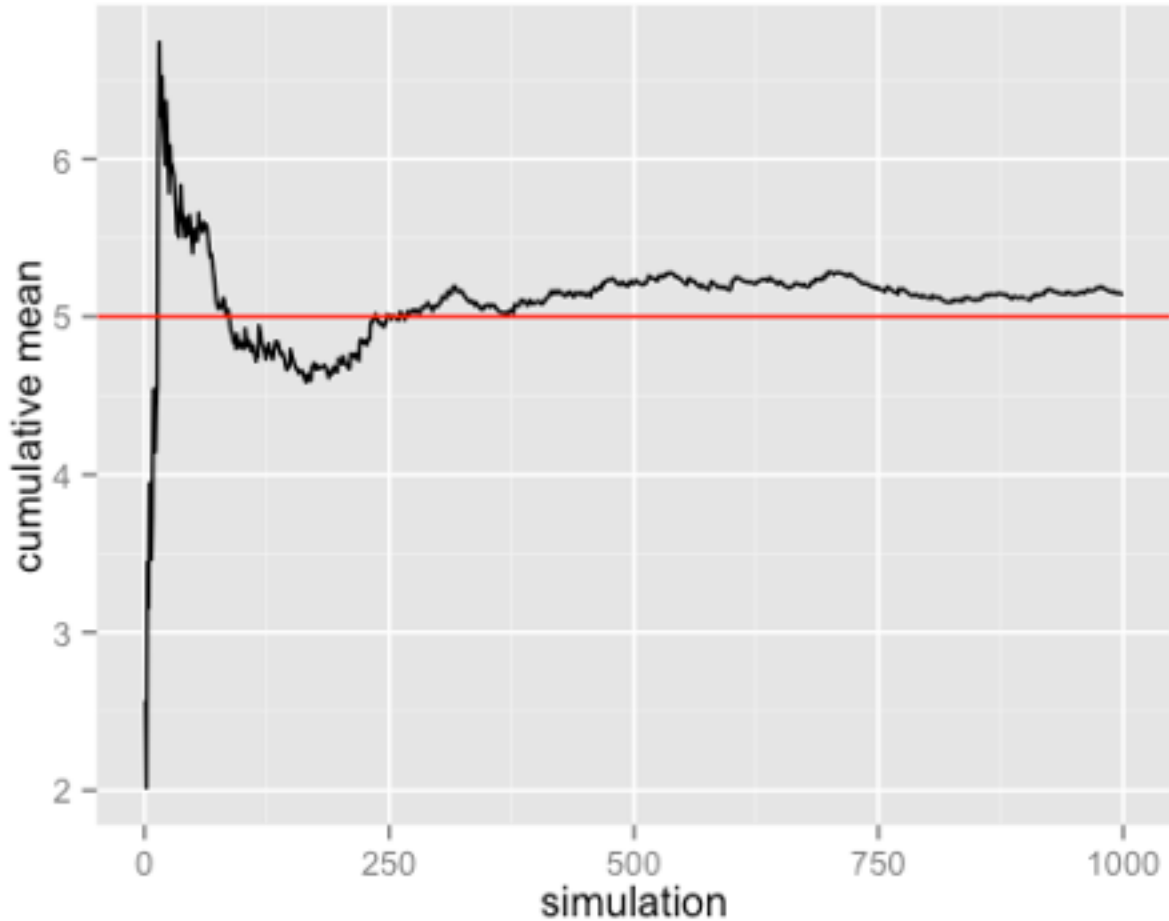
```
# approximation of sample variance 'asv' from 'asm' since mean = sd
and var = sd^2
asv <- asm^2
# theoretical sample variance 'sv' for exponential distribution = (1 /
lambda)^2
sv <- (1 / lambda)^2
```

- approximation of sample variance = 25.1638291

- theoretical sample variance = 25

2.3 Show that the distribution is approximately normal

Distribution of sample means:

```
# plot of means
hist(data_mns, prob = TRUE, main = "", xlab = "sample means")
abline(v=sm, col = "Red", lwd = 2)
lines(density(data_mns))
```
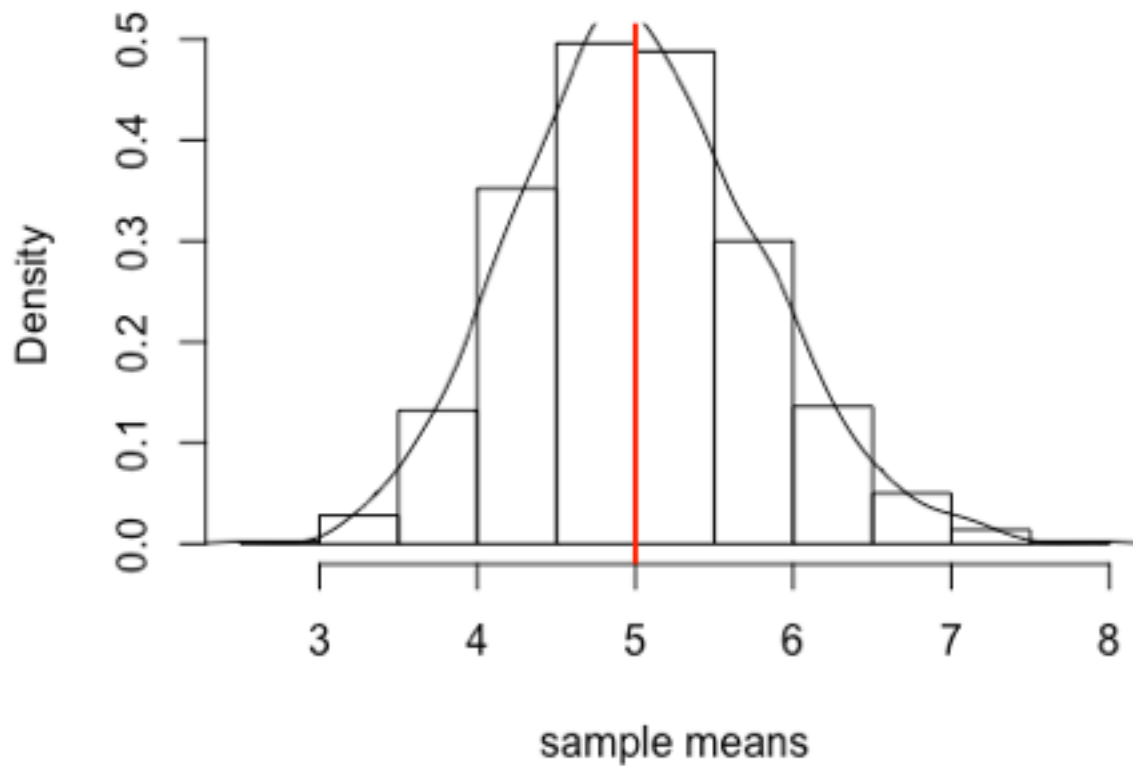
3. RESULTS

3.1 Estimated vs. Theoretical sample mean

The estimate value of the mean of the 1000 sample means of 40 exponentials is very close to the theoretical sample mean of the distribution.

3.2 Estimated vs. Theoretical sample variance

The variability of the sample means decreases as the number of simulation increases and its value gets close to the true population value as shown in the first graph. This illustrates the Law of Large Numbers (LLN) that states: "the sample mean of iid samples is consistent for the population mean."

The LLN also states: "the sample variance and the sample standard deviation of iid random variables are consistent as well." Therefore, we can infere, from the variability of estimated sample mean, that estimated sample variance will be close to the theoretical variance of the distribution.

3.3 Normality of the exponential distribution

The Central Limit Theorem (CLT) states: "the distribution of averages of iid variables becomes that of a standard normal as the sample size increases." In our simulation, the second graph shows that the distribution of the means of 40 exponentials is "quasi" normal, centered around the true sample mean (5).