Statistical Inference Course Project

Part 1: Simulation Exercises

Omer Kara

Friday, January 15, 2016

Source code for this entire report can be found here.

Overview

The exponential distribution (a.k.a. negative exponential distribution) is the probability distribution that describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate.[1] It is a special case of the gamma distribution. The exponential distribution can be simulated in R with `rexp(n, lambda)` where `lambda` such that $\lambda > 0$ is the parameter of the distribution, often called the rate parameter. The distribution is supported on the interval $[0, \infty)$. If a random variable X has this distribution, we write $X \sim Exp(\lambda)$. Note that the mean and standard deviation of exponential distribution is $1/\lambda$. For more information about the moments of exponential distribution please see Exponential Distribution Wiki Page.

For this simulation exercie, $\lambda = 0.2$ is given. In this simulation, we investigate the distribution of averages of 40 observations sampled from exponential distribution with $\lambda = 0.2$ with 1000 simulations.

Simulation

```
set.seed(1)
lambda <- 0.2
obs <- 40
sim <- 1000
data <- data.frame(means = apply(matrix(rexp(sim * obs, rate = lambda), sim), 1, mean))
```

Theoretical vs. Empirical Distribution

The central limit theorem (CLT) states that averages of sufficiently larger number of independently and identically distributed observations, each with a well-defined expected value and well-defined variance, will be asymptotically normally distributed, regardless of the underlying distribution. Therefore, following the theory, we are expecting to see our simulation data is distributed as $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$. More specifically, in our case the theoretical distribution is as follows $\overline{X} \sim \mathcal{N}(\lambda^{-1}, \lambda^{-2}/n)$.

The following results clearly shows that our simulation coincides with the theosry very well since the moments of empirical distribution is very close to the theoretical counterparts.

```
## Theoretical vs. Empirical mean.
mu <- c(1/lambda, mean(data$means))
mu
```

```
## [1] 5.000000 4.990025
```

[1] https://en.wikipedia.org/wiki/Exponential_distribution

```r
## Theoretical vs. Empirical variance.
var <- c(1/((lambda ^ 2)*obs), var(data$means))
var
```

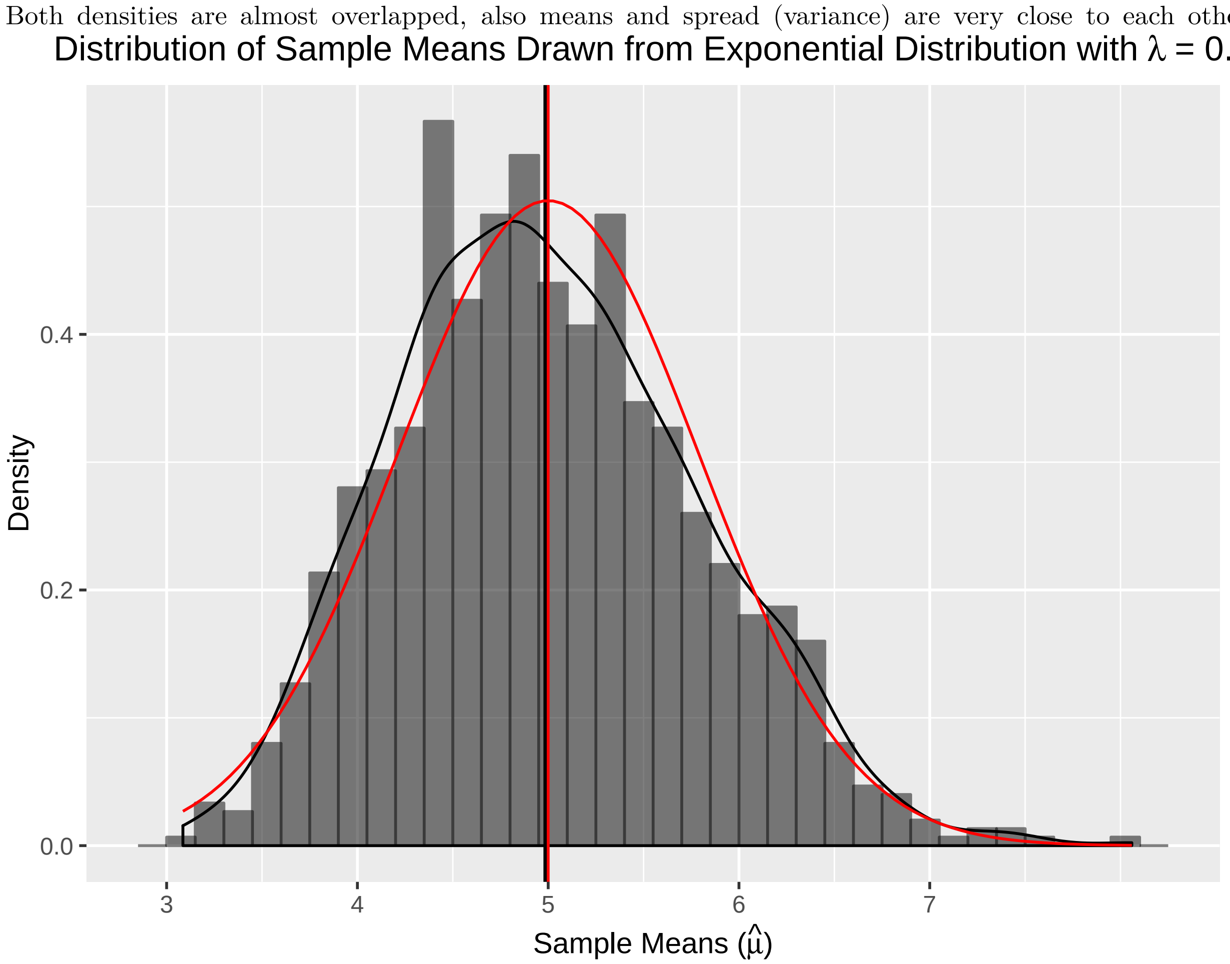
```
## [1] 0.6250000 0.6177072
```

```r
## Theoretical standard deviation vs. Empirical standard error.
sd <- c(sqrt(1/((lambda ^ 2)*obs)), sd(data$means))
sd
```

```
## [1] 0.7905694 0.7859435
```

Empirical and Theoretical Distributions

The figure below shows the theoretical distribution with `red` lines and empirical distribution with `black` lines.[2] As it is clearly shown, the empirical distribution is asymptotically normally distributed. Both densities are almost overlapped, also means and spread (variance) are very close to each other.



[2] Veritcal lines are the respective means.

Formal Normality Tests

The Shapiro-Wilk and Anderson-Darling normality test suggests that the empirical distribution is in fact normal. Both tests fail to reject the null hypothesis which indicates the data is normal.

```r
shapiro.test(data$means) ## Normality check with Shapiro-Wilk test.
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$means
## W = 0.99157, p-value = 1.759e-05
```

```r
ad.test(data$means) ## Normality check with Anderson-Darling test.
```

```
##
##  Anderson-Darling normality test
##
## data:  data$means
## A = 1.9229, p-value = 6.617e-05
```