Exponential Distribution Simulation

Di Zien LOW, dizien@gmail.com

Tuesday, December 16, 2014

Introduction

This is Part 1 of the Statistical Inference Coursework. Key objectives of this report is to document the simulation of large number of exponential distributions using R, and highlight the observed findings about the distribution of their averages. This report is structured into 3 sections:

Section 1. Problem definitions

Section 2. Simulation & findings

Section 3. Additional investigations

Section 1. Problem definitions

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. Set lambda = 0.2 for all of the simulations.

In this simulation, you will investigate the distribution of averages of 40 exponential(0.2)s. Note that you will need to do a thousand or so simulated averages of 40 exponentials.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponential(0.2)s. You should answer the following:

Q1. Show where the distribution is centered at and compare it to the theoretical center of the distribution.

Q2. Show how variable it is and compare it to the theoretical variance of the distribution.

Q3. Show that the distribution is approximately normal

Section 2. Simulation & findings

Using a step-by-step approach, the exponential distributions were simulated and analysed as follow:

2a. Set seed & run n-simulations

2b. Compare simulated vs theoretical means (Q1)

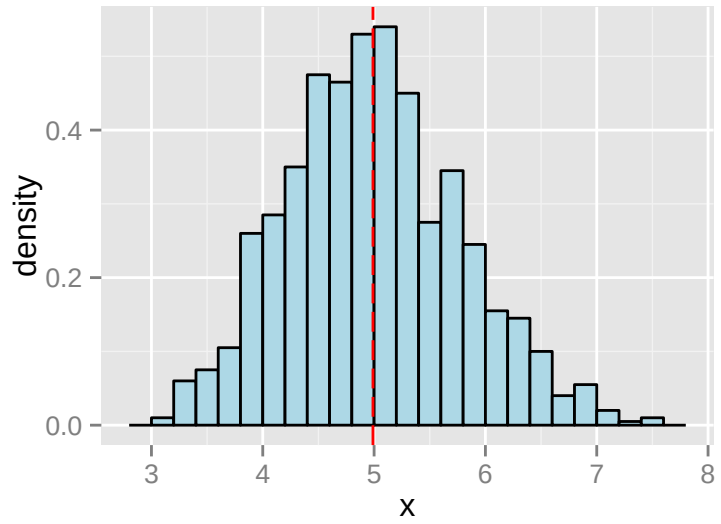
2c. Compare simulated vs theoretical variances (Q2)

2d. Compare distribution of averages vs normal distribution (Q3)

2a. Set seed & run n-simulations

```
lambda <- 0.2
n <- 40
nsim <- 1:1000
set.seed(1)
simavg <- data.frame(x = sapply(nsim, function(x) {mean(rexp(n, lambda))}))
```

```r
library(ggplot2)
## plot histogram of simulated averages
g <- ggplot (data=simavg, aes(x=x)) + geom_histogram (aes(y=..density..),
             binwidth=0.2, fill="light blue", color="black")
## plot dotted red-line representing the mean of simulated averages
g + geom_vline (xintercept = mean(simavg$x), color="red", linetype="longdash")
```



2b. Compare simulated vs theoretical means (Q1)

```r
## mean of simulated averages
mean(simavg$x)
```

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

```r
## theoretical mean of exponential distribution
1/lambda
```

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

```r
## differences between simulation and theoretical mean
paste(round(abs(mean(simavg$x)-1/lambda) / (1/lambda) *100, 3),"%")
```

```
## [1] "0.199 %"
```

Findings:

- Simulated mean (4.99) is very close to theoretical mean (5.00)
- Simulated mean only differ with theoretical mean by about 0.199%

2c. Compare simulated vs theoretical variances (Q2)

```
## variance of simulated averages
var(simavg$x)
```

```
## [1] 0.6111165
```

```
## theoretical sampling variance
1/(n * lambda^2)
```

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

```
## differences between simulation and theoretical mean
paste(round(abs(var(simavg$x)-1/(n * lambda^2))/(1/(n * lambda^2))*100, 2),"%")
```

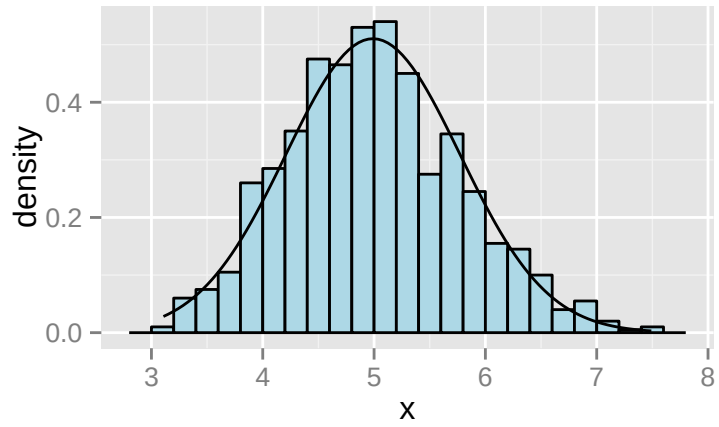
```
## [1] "2.22 %"
```

Findings:

- Simulated variance (0.611) is very close to theoretical sampling variance (0.625)
- Simulated variance only differ with theoretical sampling variance by about 2.22%

2d. Compare distribution of averages vs normal distribution (Q3)

```
library(ggplot2)
## plot histogram of simulated averages
g <- ggplot (data=simavg, aes(x=x)) + geom_histogram (aes(y=..density..),
             binwidth=0.2, fill="light blue", color="black")
## plot line representing the standard normal distribution
g + stat_function(fun=dnorm, arg=list(mean=mean(simavg$x), sd = sd(simavg$x)))
```



Findings:

- As expected, distribution of simulated averages fits quite closely with normal distribution

Section 3. Additional investigations

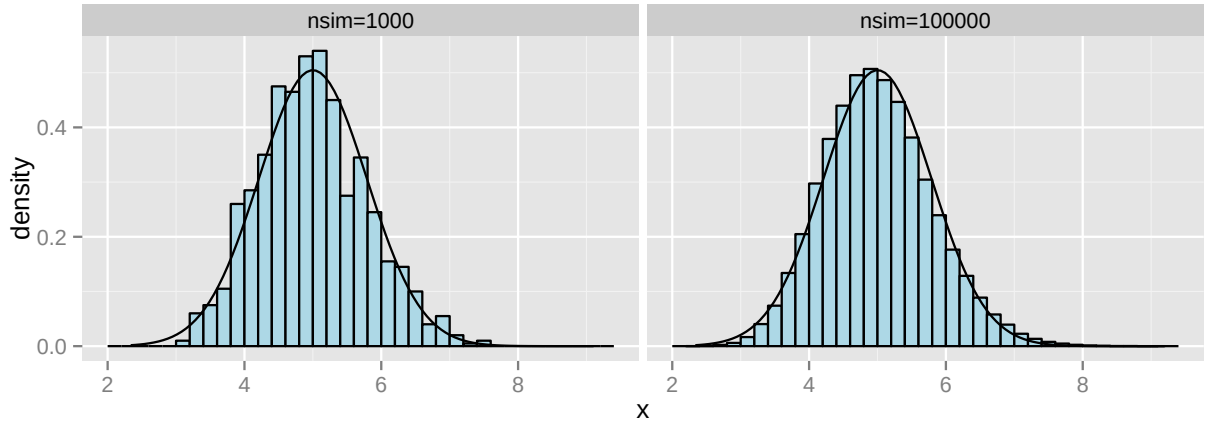
The Central Limit Theorem describes the characteristics of the "population of the means" which has been created from the means of an infinite number of random population samples of size (N), all of them drawn from a given "parent population". In our case, this parent poplulation is an exponential distribution.

Using similar approach of simulating exponential distributions, let's compare the distribution of their averages between 1000 simulations vs 100000 simulations

```
## run simulations
lambda <- 0.2
n <- 40
set.seed(1)
nsim <- 1:1000
simavg1 <- data.frame(x = sapply(nsim, function(x) {mean(rexp(n, lambda))}),
                      y= sapply(nsim, function(y) {"nsim=1000"}))
nsim <- 1:100000
simavg2 <- data.frame(x = sapply(nsim, function(x) {mean(rexp(n, lambda))}),
                      y= sapply(nsim, function(y) {"nsim=100000"}))
simavg <- rbind(simavg1, simavg2)

## histogram of simulated averages with expected distribution
g <- ggplot (data=simavg, aes(x=x)) + geom_histogram (aes(y=..density..),
             binwidth=0.2, fill="light blue", color="black")
g1 <-g + stat_function(fun=dnorm, arg=list(mean=5, sd = sqrt(0.625)))

## plot the two histograms side by side
g1 + facet_wrap(~y, ncol=2)
```



Regardless of the distribution of the parent population, the plots above shows that:

- The mean of the population of means is always equal to the mean of the parent population from which the population samples were drawn.
- The standard deviation of the population of means is always equal to the standard deviation of the parent population divided by the square root of the sample size (N).
- The distribution of means will increasingly approximate a normal distribution as the size N of samples increases. Amazing!!