Analysis of the exponential distribution in R

RDSN
20 October 2015

1. Overview

In this project, we investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution is simulated in R with the function `rexp(n, lambda)`, with `lambda` being the rate parameter, `1/lambda` being the mean of the distribution and `1/lambda` being also the standard deviation of the distribution. We will run here several simulations to compare the Sample Mean with the Theoritical Mean, the Sample Variance with the Theoritical Variance, and to show that the distribution is approximately normal.

2. Simulations

We will run here 1000 simulations each time. For each simulation, we set the parameter `lambda = 0.2`. And we will investigate the averages of 40 exponentials.

We first here set the parameters we are going to use through all the simulations.

```
lambda <- .2
nosim <- 1000
n <- 40
```

2.1 Sample Mean versus Theoritical Mean

We want to show here that the expected value of the Sample Mean is the Population Mean that it's trying to estimate. Here, we take 1000 values from the same population, a population with an exponential distribution, with a rate lambda of .2. The mean of this population is so 1/lambda = 1/.2 = 5. To show that the expected value of the Sample Mean is the Population Mean (5), we simulate 1000 averages of 40 exponentials from the same population.

```
library(ggplot2)
set.seed(3)

# Here we create a matrix with the first 1000 rows corresponding to the random exponentials, and the following
1000 rows corresponding to the averages of 40 exponentials.
dat1 <- data.frame(
    x = c(rexp(nosim,lambda), apply(matrix(rexp(nosim * n,lambda), nosim), 1, mean)),
    what = factor(rep(c("Obs", "Averages"), c(nosim, nosim)))
    )

# Then we plot the density function for the 2 sets, observation (random exponentials) and averages(averages of
40 exponentials)
g <- ggplot(dat1, aes(x = x, fill = what)) + geom_density(size = .5, alpha = .2)
g <- g + geom_vline(x = 1/lambda, size = 1, col = "red")
g
```

See Appendix – Figure 1 for the plot

What we can see on the figure above is that the averages of 40 exponentials (red figure), is centered around the mean of the population, 1/lambda = 5 (red line). What that means is that the expected value of the Sample Mean is the Population Mean, here 1/lambda = 5.

Let's have a look at the mean of the averages of the 40 exponentials shown in the plot above, and compare it to the theoritical mean.

```
empirical <- round(mean(dat1$x),3)
theoritical <- 1/lambda
```

The empirical mean is 4.953 and the theoritical mean is 5.

2.2 Sample Variance versus Theoritical Variance

The variance of a random variable is a measure of spread.

First have a look at the distribution with increasing variance to see how variable is the sample depending on the variance.

```
set.seed(3)

# First we create a vector for x values from -5 to 5 by 0.01
xvals <- seq(-5, 5, by = .01)
# Then we create a dataframe with values from the density function of the exponential distribution, dexp, for 4
different rates, to see the variability of the distribution depending on the variance( 1/ rate^2). We aslo add
the x values to this dataframe and the variance value so that to make a separation into the plot according to t
his variance.
dat2 <- data.frame(
    y = c(
        dexp(xvals, rate = 0.1),
        dexp(xvals, rate = 0.5),
        dexp(xvals, rate = 1),
        dexp(xvals, rate = 2)
    ),
    x = rep(xvals, 4),
    variance = factor(rep(c(100,4,1,0.25), rep(length(xvals), 4)))
)
ggplot(dat2, aes(x = x, y = y, color = variance)) + geom_line(size = 1)
```

See Appendix – Figure 2 for the plot

Here, we take 1000 values from the same population, a population with an exponential distribution, with a rate lambda of .2. The variance of this population is so 1/lambda^2 = 1/.2^2 = 25. To show that the expected value of the Sample variance is the Population variance (25), we simulate 1000 variances of 10, 40, and 80 exponentials from the same population.

```
set.seed(3)

# We create a dataframe which contains the variances of 10 exponentials for the first 1000 rows, then the varia
nces of 40 exponentials for the next 1000 rows, and finally the variances of 80 exponentials for the last 1000
rows. We associate those values with the factor n which enable to make a separation into the plot according to
the number of exponentials averaged.
dat3 <- data.frame(
    x = c(apply(matrix(rexp(nosim * 10,lambda), nosim), 1, var),
          apply(matrix(rexp(nosim * 40,lambda), nosim), 1, var),
          apply(matrix(rexp(nosim * 80,lambda), nosim), 1, var)),
    n = factor(rep(c("10", "40", "80"), c(nosim, nosim, nosim)))
    )
# We plot the density function for the 3 values of variances of exponentials, and we plot a line corresponding
to the variance of the population.
ggplot(dat3, aes(x = x, fill = n)) + geom_density(size = .5, alpha = .2) + geom_vline(xintercept = 1/(lambda^2)
, size = 1, col = "red")
```

See Appendix – Figure 3 for the plot

What we can see on the figure above is that the variance varies depending on the sample size. The larger the size, the more the sample variance become narrower from the population variance, which is here represented by the red line, for a value of 25 (1/lambda^2).

The variance of sample mean is σ^2/n

Let's try a simulation to see that.

```
set.seed(3)

empirical2 <- round(var(apply(matrix(rexp(nosim * n, lambda), nosim), 1, mean)),3)
theoritical2 <- 1 / ((lambda^2) * n)
empirical2

## [1] 0.626

theoritical2

## [1] 0.625
```

So the variance of the sample set is 0.626 and the theoritical variance is 0.625.

2.3 Distribution

Let's compare the distribution of:

- a large collection of random exponentials
- a large collection of averages of 40 exponentials

```
set.seed(3)

# We create a dataframe containing 1000 random exponentials for the first 1000 rows and the averages of 40 expo
nentials for the next 1000 rows.
dat4 <- data.frame(
    x = c(rexp(nosim,lambda), apply(matrix(rexp(nosim * n,lambda), nosim), 1, mean)),
    what = factor(rep(c("Obs", "Averages"), c(nosim, nosim)))
    )
# we plot the histogram with the distinction between the 2 sets of values, and we plot the density function so
that we can see the correspondance with the normal distribution
g <- ggplot(dat4, aes(x = x))
g <- g + geom_histogram(aes(y = ..density..), binwidth = .5) + geom_density(col = "red", size = 1)
g <- g + facet_grid(.~ what)
g <- g + coord_cartesian(xlim = c(0, 10))
g
```

See Appendix – Figure 4 for the plot

We can clearly see that the distribution of the large collection of averages of 40 exponentials (on the left) is narrower to a normal distribution than the distribution from the collection of exponentials (on the right).

Appendix

Figure 1 - density function for the 2 sets,

observation (random exponentials) and averages (averages of 40 exponentials)

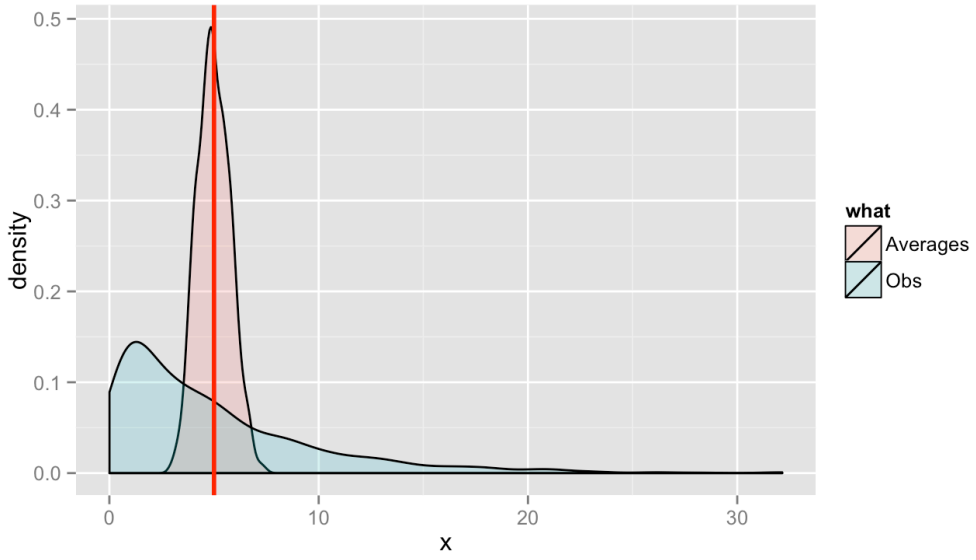

Figure 2 - density function of the exponential distribution,

for 4 different variances

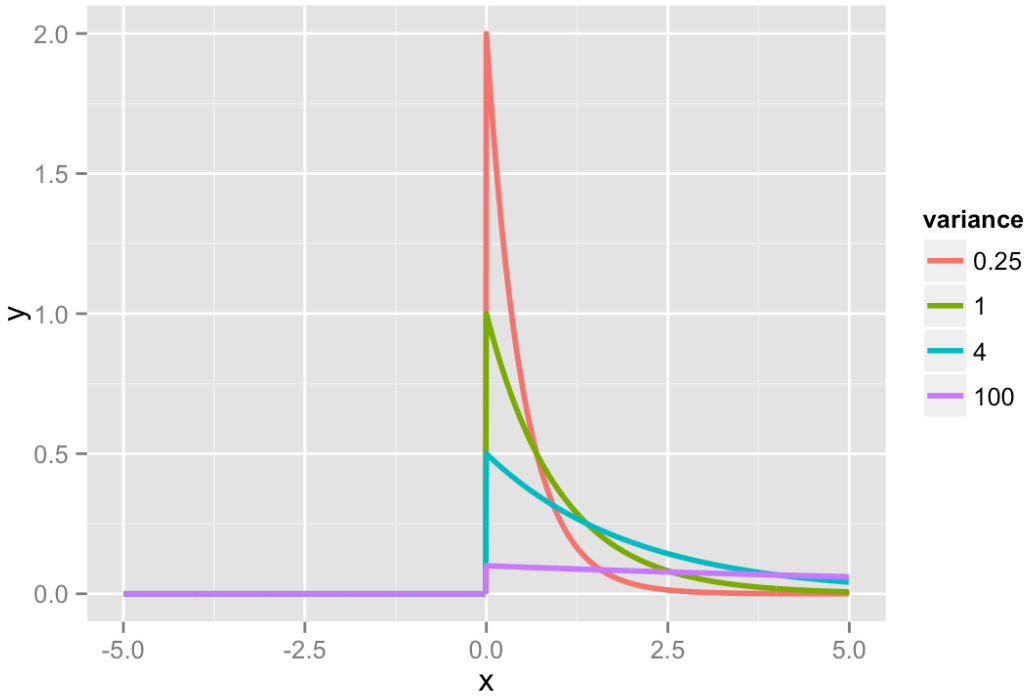

Figure 3 - density function for the 3 values of variances of exponentials compared to the variance of the population

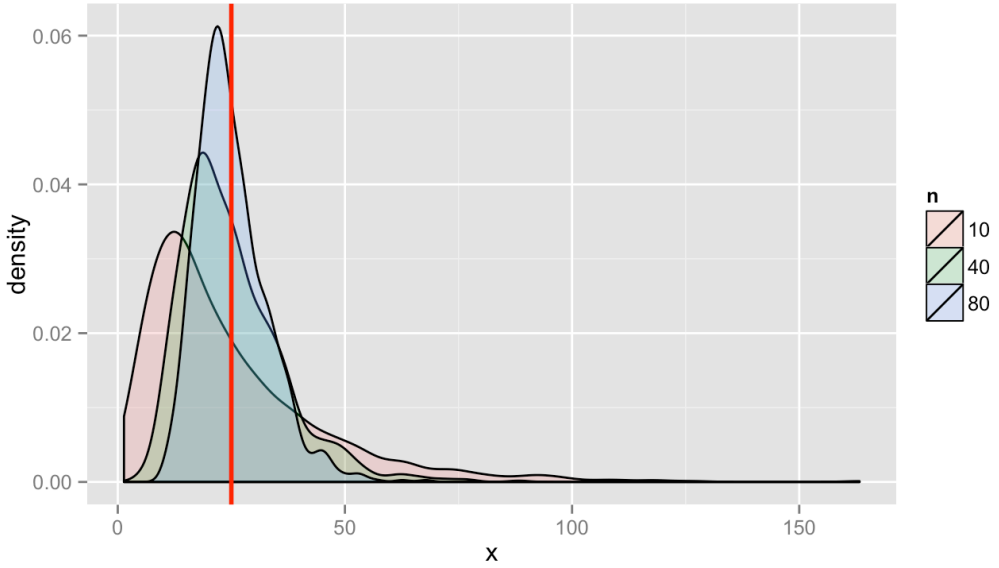

Figure 4 - histogram of observations (random exponentials) and averages (averages of 40 exponentials), and density function so that we can see the correspondance with the normal distribution