Investigation of the exponential distribution in R

Stephanie

February 2016

Contents

Overview

In this project we will investigate the exponential distribution in R and compare it with the Central Limit Theorem. 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 1/lambda. We will set lambda = 0.2 for all of the simulations. We will investigate the distribution of averages of 40 exponentials.

We will show:

- the sample mean and compare it to the theoretical mean of the distribution
- how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
- that the distribution is approximately normal.

Simulations

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

We first set the parameters that we are going to use through all the simulations:

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

Sample mean versus Theoritical mean

We will see that the expected value of the sample mean is the population mean that it's trying to estimate. We take 1000 values from the same population, a population with an exponential distribution and 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 sean is the population mean (5), we simulate 1000 averages of 40 exponentials from the same population.

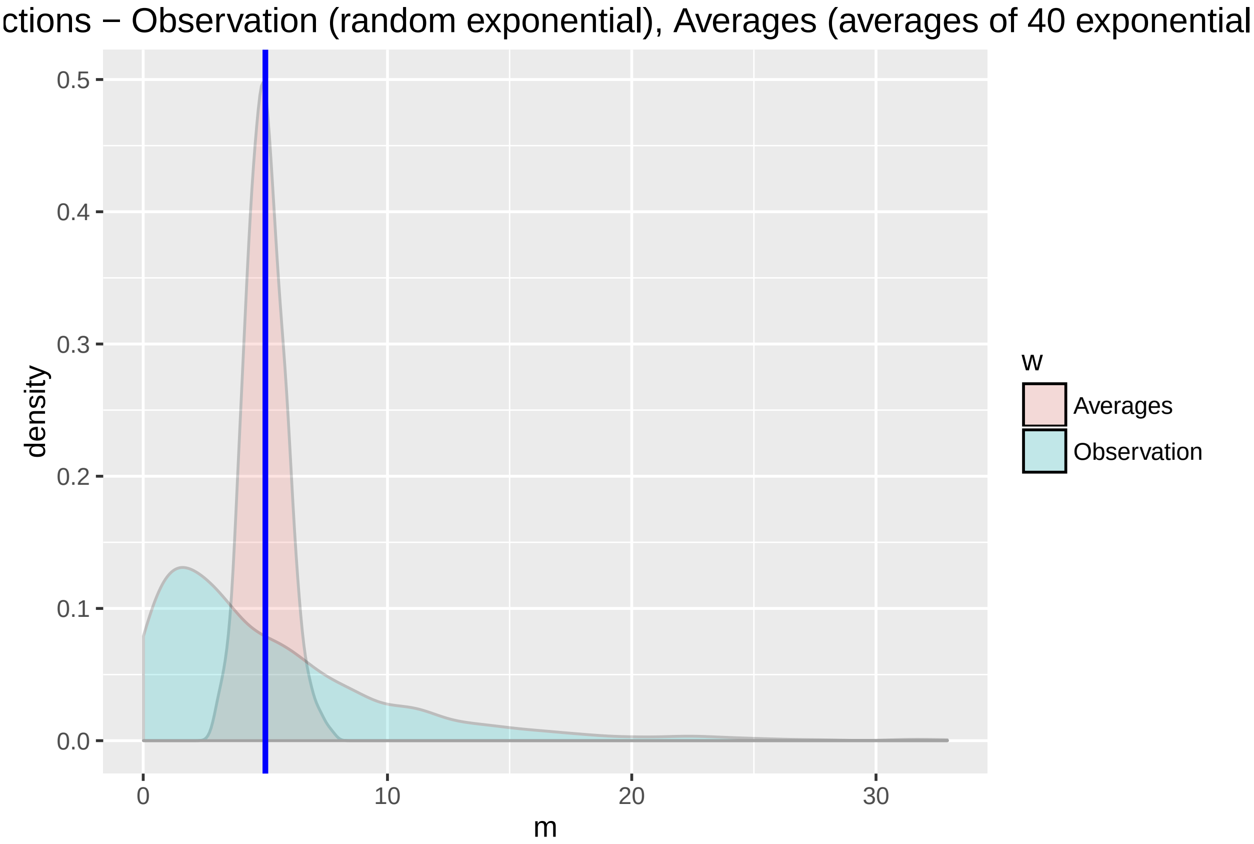
```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.3
```

```
set.seed(0)

# We create a data frame with the first 1000 rows corresponding to the random exponentials, and the fol
data1 <- data.frame(
    m = c(rexp(sim,lambda), apply(matrix(rexp(sim * n,lambda), sim), 1, mean)),
    w = factor(rep(c("Observation", "Averages"), c(sim, sim)))
    )

# We plot the density function for the 2 sets, observation (random exponentials) and average (averages
plot1 <- ggplot(data1, aes(x = m, fill = w)) + geom_density(size = .5, alpha = .2)
plot1 <- plot1 + geom_vline(xintercept = 1/lambda, size = 1, col = "blue")
plot1 <- plot1 + labs(title = "Density functions - Observation (random exponential), Averages (averages
plot1
```



We can see on the figure that the averages of 40 exponentials (red figure), is centered around the mean of the population, 1/lambda = 5 (blue line). 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(data1$x),3)
```

```
## Warning in mean.default(data1$x): argument is not numeric or logical:
## returning NA
```

```
theoritical <- 1/lambda
```

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

Sample variance versus Theoritical variance

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

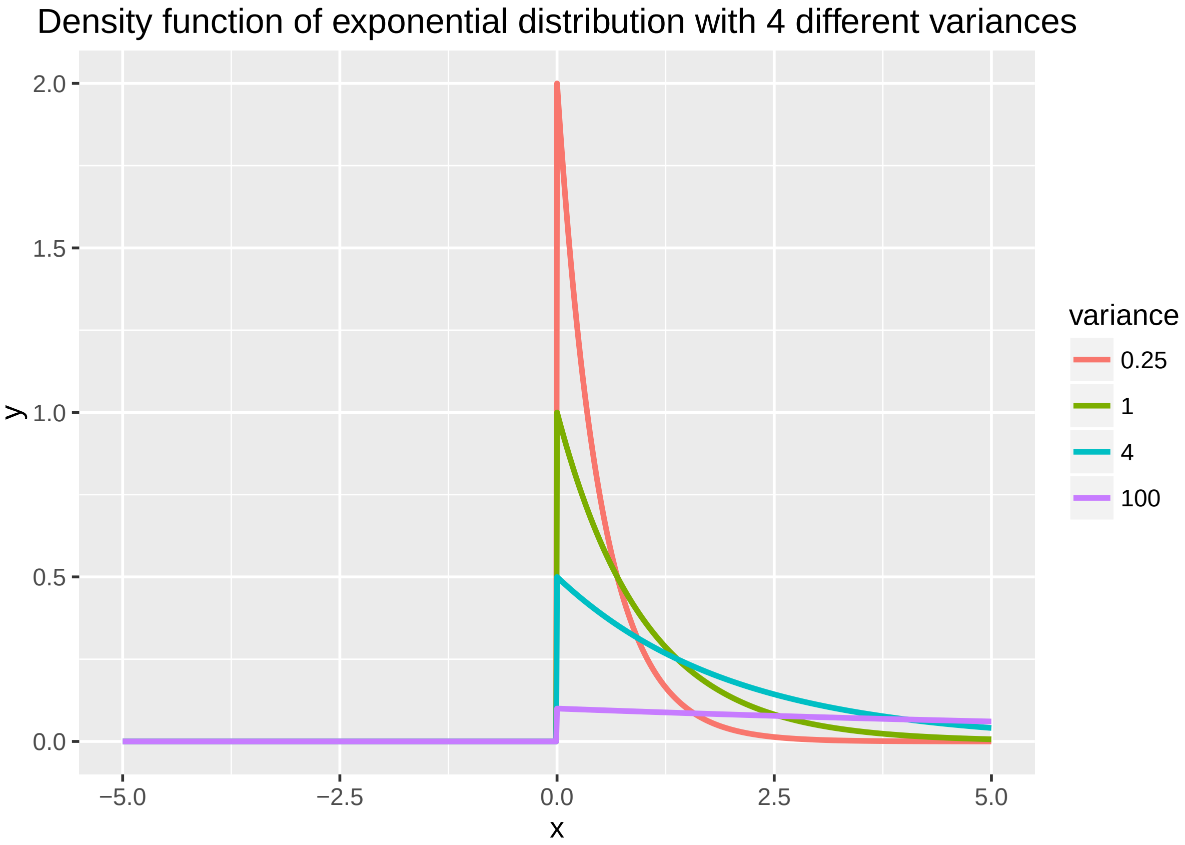
Let's first have a look at the distribution with increasing variance to see how variable is the sample depending on the variance.

```
set.seed(0)

# We create a vector for x values from -5 to 5 by 0.01
vect <- seq(-5, 5, by = .01)

# We create a dataframe with values from the density function of the exponential distribution, dexp, fo
data2 <- data.frame(
    y = c(
        dexp(vect, rate = 0.1),
        dexp(vect, rate = 0.5),
        dexp(vect, rate = 1),
        dexp(vect, rate = 2)
    ),
    x = rep(vect, 4),
    variance = factor(rep(c(100,4,1,0.25), rep(length(vect), 4)))
)

ggplot(data2, aes(x = x, y = y, color = variance)) + geom_line(size = 1) + labs(title = "Density functi
```

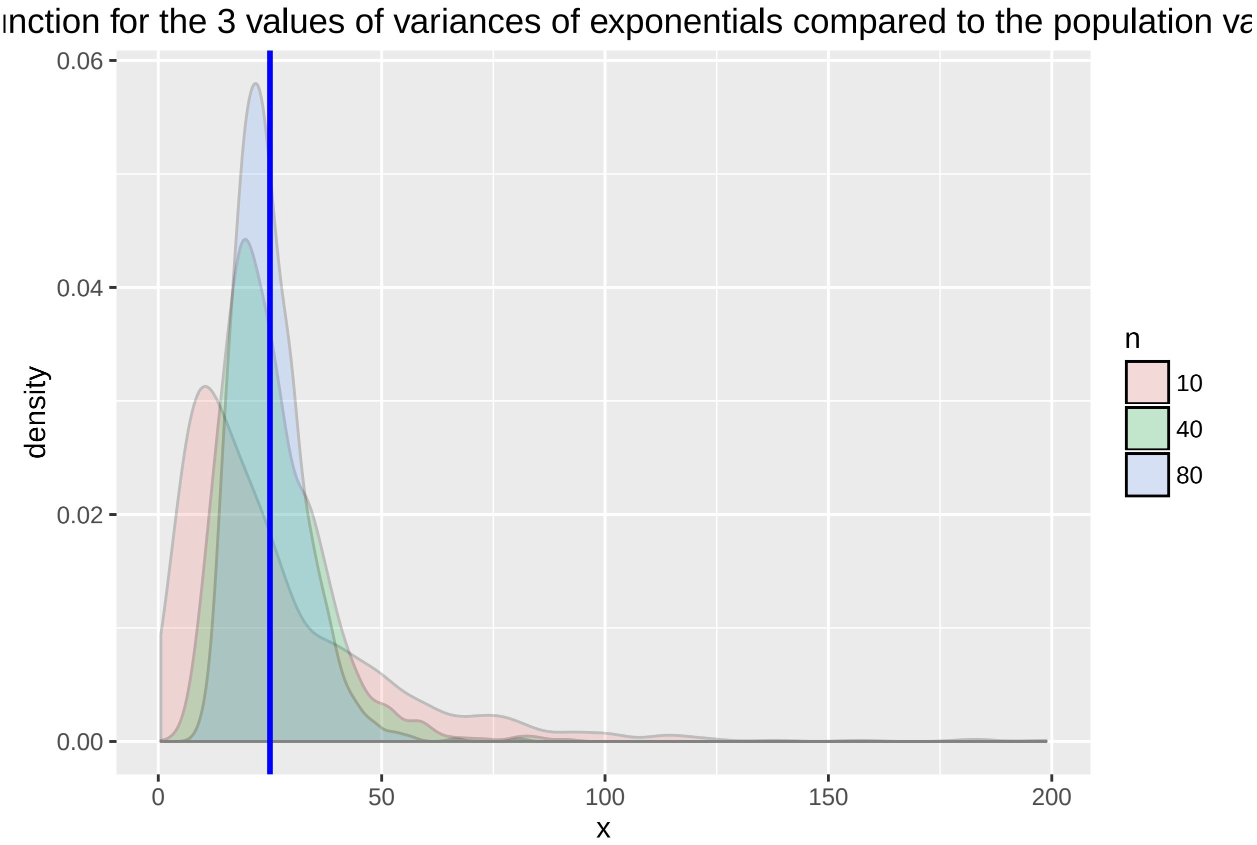

Density function of exponential distribution with 4 different variances

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/lambdaˆ2 = 1/.2ˆ2 = 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(0)

# We create a dataframe which contains the variances of 10 exponentials for the first 1000 rows, then t
data3 <- data.frame(
    x = c(apply(matrix(rexp(sim * 10,lambda), sim), 1, var),
          apply(matrix(rexp(sim * 40,lambda), sim), 1, var),
          apply(matrix(rexp(sim * 80,lambda), sim), 1, var)),
    n = factor(rep(c("10", "40", "80"), c(sim, sim, sim)))
    )

# We plot the density function for the 3 values of exponentials averaged, and we plot a line correspond
ggplot(data3, aes(x = x, fill = n)) + geom_density(size = .5, alpha = .2) + geom_vline(xintercept = 1/(
```



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 blue line, for a value of 25 (1/lambdaˆ2).

The variance of sample mean is σ^2/n

Let's try a simulation to see that.

```
set.seed(0)

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

```
## [1] 0.664
```

```
theoritical2
```

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

The variance of the sample set is 0.664 and the theoritical variance is 0.625.

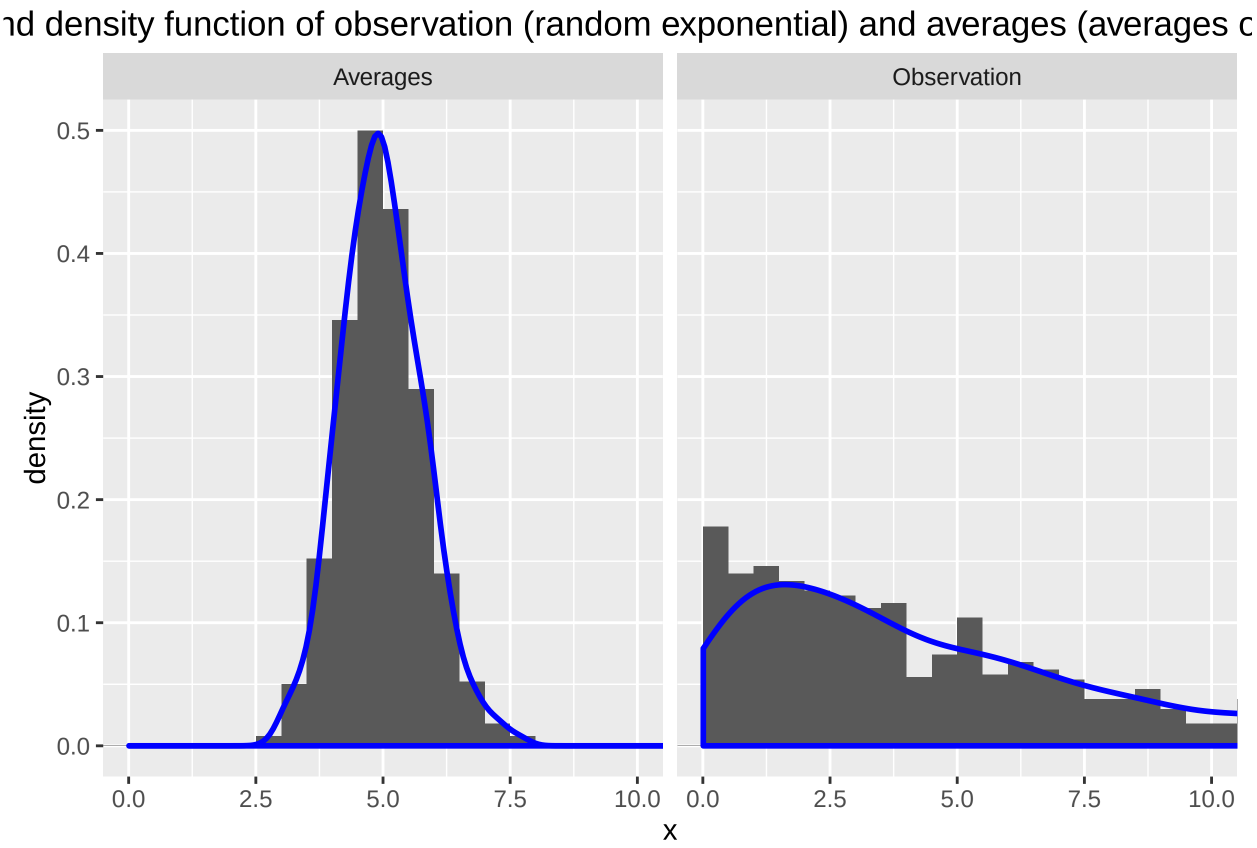
Distribution

Let's compare the distribution of:

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

```
set.seed(0)

# We create a dataframe containing 1000 random exponentials for the first 1000 rows and the averages of
data4 <- data.frame(
    x = c(rexp(sim,lambda), apply(matrix(rexp(sim * n,lambda), sim), 1, mean)),
    what = factor(rep(c("Observation", "Averages"), c(sim, sim)))
    )
# we plot the histogram with the distinction between the 2 sets of values, and we plot the density func
plot2 <- ggplot(data4, aes(x = x))
plot2 <- plot2 + geom_histogram(aes(y = ..density..), binwidth = .5) + geom_density(col = "blue", size =
plot2 <- plot2 + facet_grid(.~ what)
plot2 <- plot2 + coord_cartesian(xlim = c(0, 10)) + labs(title = "Histogram and density function of obs
plot2
```



We can 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). The distribution is approximately normal.