A simulation exercise

Katica Ristic

July 2, 2017

Overview

Simulation of thousand averages of 40 **exponentials** in R will be compared with **Central Limit Theorem**, which indicate that distribution of iid variables becomes that of a standard normal as the sample size increases. Also the sample mean and the sample variance will be compared with the theoretical ones. Rate parameter, λ, for the exponentials is set to 0.2 for all the simulations.

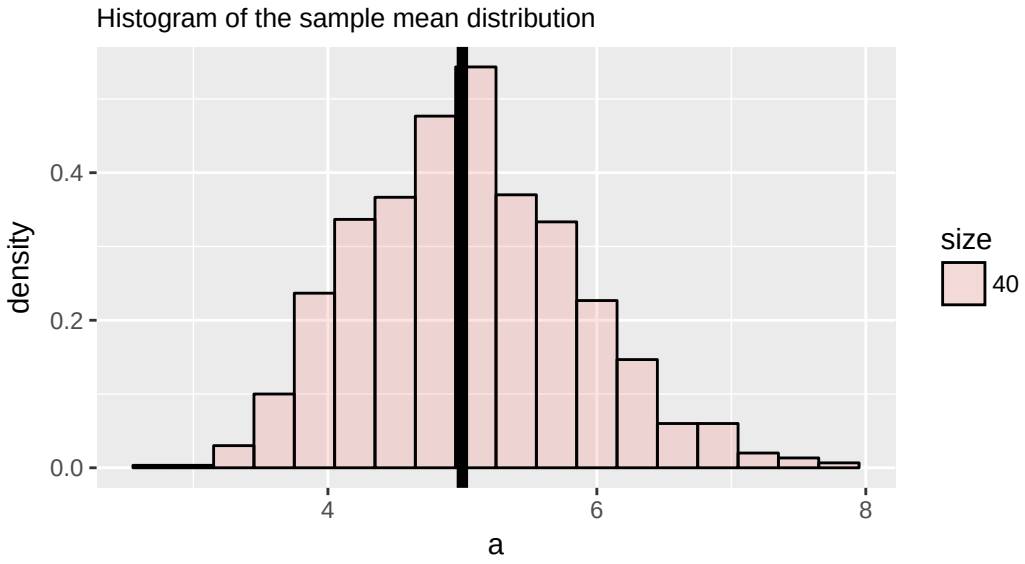
Simulation

Let us simulate thousand times 40 exponentials and put it in matrix `m`. Then calculate the mean of each row and put it in the vector `a`. Now we can calculate the mean and standard error of the vector with the means, `a`, and compare it with the theoretical mean and the theoretical standard deviation.

```
set.seed(10)
nosim<-1000
n<-40
lambda<-.2
m<-matrix(rexp(nosim * n,lambda), nosim)
a<-apply(m, 1, mean)
data.frame("sample"=round(c(mean(a),sd(a)),2),
           "theoretical"=round(c(5,5/sqrt(n)),2),
           row.names = c("mean","se"))
```

```
##      sample theoretical
## mean   5.05        5.00
## se     0.81        0.79
```

As you can see, the theoretical and empirical results are pretty close, right?



Here is a histogram of our sample, it is very Gaussian, like a bell curve, centered at the black vertical line. This line represents the theoretical mean of our sample, $1/\lambda = 5$. From the theoretical data we can say with 95% confidence that the sample mean is between 3.45 and 6.55. So if we set the null hypotesis to be $H_0 : \mu = 5$ versus alternative hypothesis, for example, $H_a : \mu > 5$ then the calculated sample mean 5.05 is in the confidence interval so we will not reject the null hypotesis.

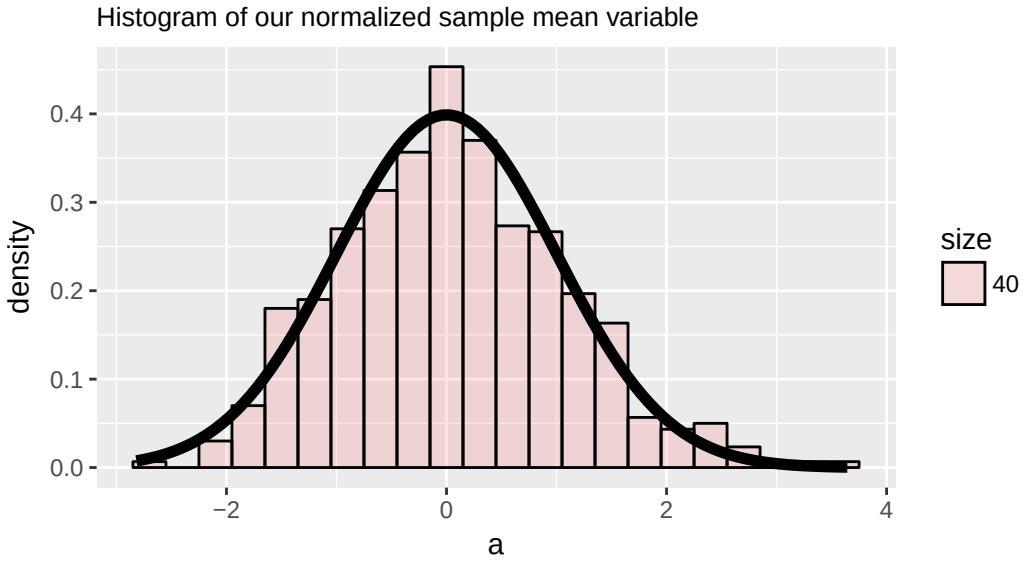
Distribution

We will now use Central Limit Theorem to confirm that the distribution of sample means is standard normal. Let us simulate once again our sample, but this time we will normalized it with the function `CTfunc` (by taking the mean of each row, subtracting off $\mu = 1/\lambda = 5$ and dividing it by $(1/\lambda)/\sqrt{n} = 5/\sqrt{40}$, since $1/\lambda$ is the standard deviation as well.)

```
set.seed(10)
nosim<-1000
CTfunc <- function(x, n) sqrt(n) * (mean(x) - 5) / 5
a<-apply(matrix(rexp(nosim * 40,.2), nosim), 1, CTfunc, 40)
round(c("mean"=mean(a),"se"=sd(a)),2)
```

```
## mean   se
## 0.06 1.02
```

Our calculation is prrety close to $N(0, 1)$. Let us confirm with a histogram as well. The black bell curve line reprezents standard normal and as you can see, the histogram is very close to it.



Summary notes

- We simulated thousand averages of 40 exponentials in R
- calculated the mean and the standard error of the variable of averages
- Theoretical and empirical results were compared
- Two histograms showed that the variable of averages is normaly distributed

Appendix

Code in R for the first histogram:

```
library(ggplot2)

set.seed(10)
nosim <- 1000

dat <- data.frame( a = apply(matrix(rexp(nosim * 40,.2), nosim), 1, mean),
                   size = factor(rep(40,nosim)))

g <- ggplot(dat, aes(x = a, fill = size))
     + geom_histogram(alpha = .20, binwidth=.3, colour = "black", aes(y = ..density..))
g <- g  + geom_vline(xintercept = 5, size = 2)
g <- g + labs(subtitle = "Histogram of the sample mean distribution")
g
```

Code in R for the second histogram:

```
library(ggplot2)

set.seed(10)
nosim <- 1000

CTfunc <- function(x, n) sqrt(n) * (mean(x) - 5) / 5
dat <- data.frame( a = apply(matrix(rexp(nosim * 40,.2), nosim), 1, CTfunc, 40),
                   size = factor(rep(40,nosim)))

g <- ggplot(dat, aes(x = a, fill = size))
     + geom_histogram(alpha = .20, binwidth=.3, colour = "black", aes(y = ..density..))
g <- g + stat_function(fun = dnorm, size = 2)
g <- g + labs(subtitle = "Histogram of our normalized sample mean variable")
g
```