Analysis of the exponential distribution and application of the Central Limit Theorem via simulation experiment

Elena Civati

Contents

Report

Synopsis

The aim of this simulation is to investigate the properties of the exponential distribution with a given rate. We'll show that mean, variance and other parameters computed from a large collection of random exponential converge on the theoretical values.
We'll also apply the Central Limit Theorem (CTL), calculating sample means from 1000 random samples of a given size (n=40) taken from the exponential distribution; as expected, we'll see that the CTL applies very well to our simulated data.

The exponential distribution

The exponential distribution is represented by the following density equation that we'll denote with γ:

$$\gamma : f(x) = \lambda e^{-\lambda x}, x \geq 0$$

It describes the time for a continuous process to change state and, under the assumption of a constant rate, it's very useful for modeling variables like time until a radioactive particle decays. See Wikipedia for more details.
Formulas for mean, variance and median can be obtained mathematically from γ:

$$E[X] = \frac{1}{\gamma} \qquad Var[X] = \frac{1}{\lambda^2} \qquad m[X] = \frac{ln(2)}{\lambda}$$

For this analysis, I generated 1000 values from an exponential distribution with rate λ=0.2.
In Table 1 are shown the population parameters, calculated with the above formulas, and the corresponding sample parameters, obtained from the simulated data (*see Appendix for code*).

Table 1: Population parameters and sample statistics for a random sample of size=1000

	Population.Value	Sample.Value
Mean	μ= 5	$\bar{X}$= 5.19
Variance	σ^2= 25	s^2= 25.76
Median	m= 3.47	$m[X]$= 3.89

Here is a histogram showing the relative frequencies of those random variables, along with their mean and median. The green line represents the population density function. We can see that the distribution is decreasing and right skewed, with mean larger than median (*see Appendix for code*).

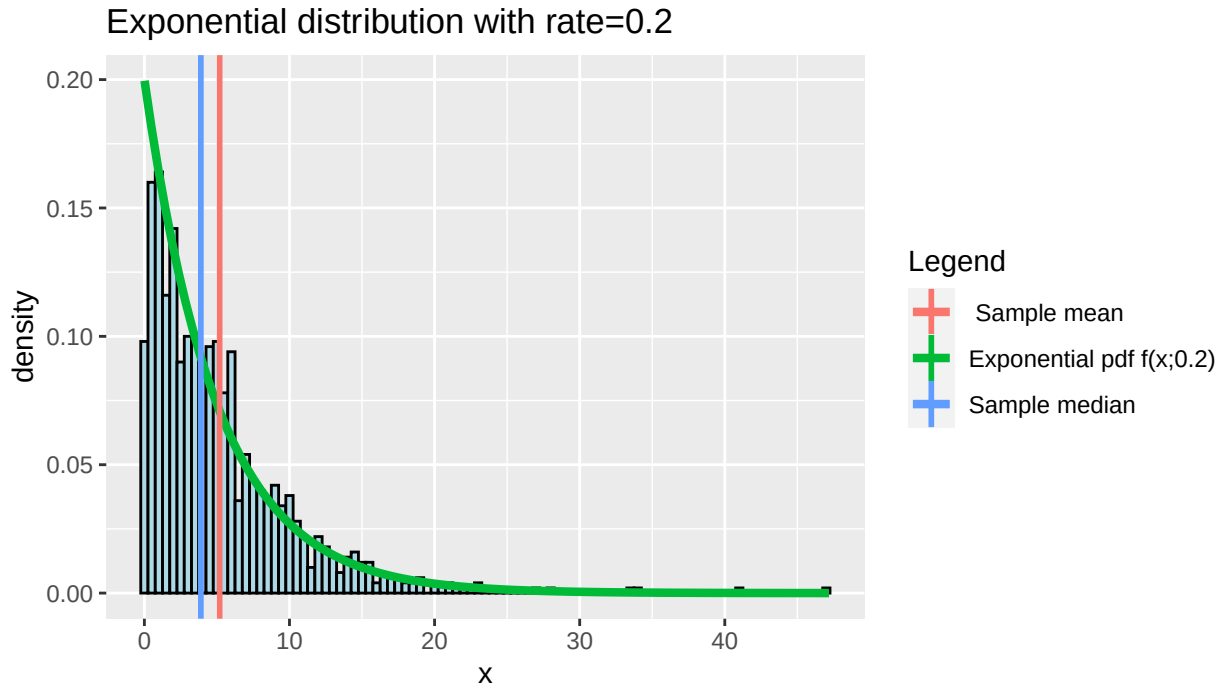

Figure 1: Plot 1

Application of Central Limit Theorem

The CTL states that, regardless of the underlying distribution of the variables, the distribution of sample means calculated from repeated samples of size n converges to a normal distribution with mean=μ and variance=$\frac{\sigma^2}{n}$ as n increases. To confirm that, we generated 1000 random exponentials of size $n = 40$ and for each we stored the average value. Using those averages as data, we calculated sample mean, median and variance and compared them, in Table 2, with the expected values of $X \sim N(\frac{1}{\lambda}, \frac{\sigma^2}{n})$ (recall that for a normal distribution, the median equals the mean) (*see Appendix for code*). We can see that the mean value is now more similar than before to the theoretical one.

Table 2: Population parameters and $\bar{X}$ statistics for 1000 sample of size 40

	Population.Value	Sample.Value
Mean	μ= 5	$\bar{X}$= 4.98
Variance	σ^2= 0.625	s^2= 0.57
Median	m= 5	$m[\bar{X}]$= 4.94

Finally, Plot 2 shows a histogram of sample means (frequencies on y axis are expressed as a proportion) together with the probability density function of the normal distribution centered at 5 with standard deviation $\frac{5}{\sqrt{40}}$ (*see Appendix for code*). We can see that the bins fit the curve pretty well (they would fit even better by increasing the sample size, according to CTL).
In particular, vertical line representing the average of samples means is very close to the population expected value.

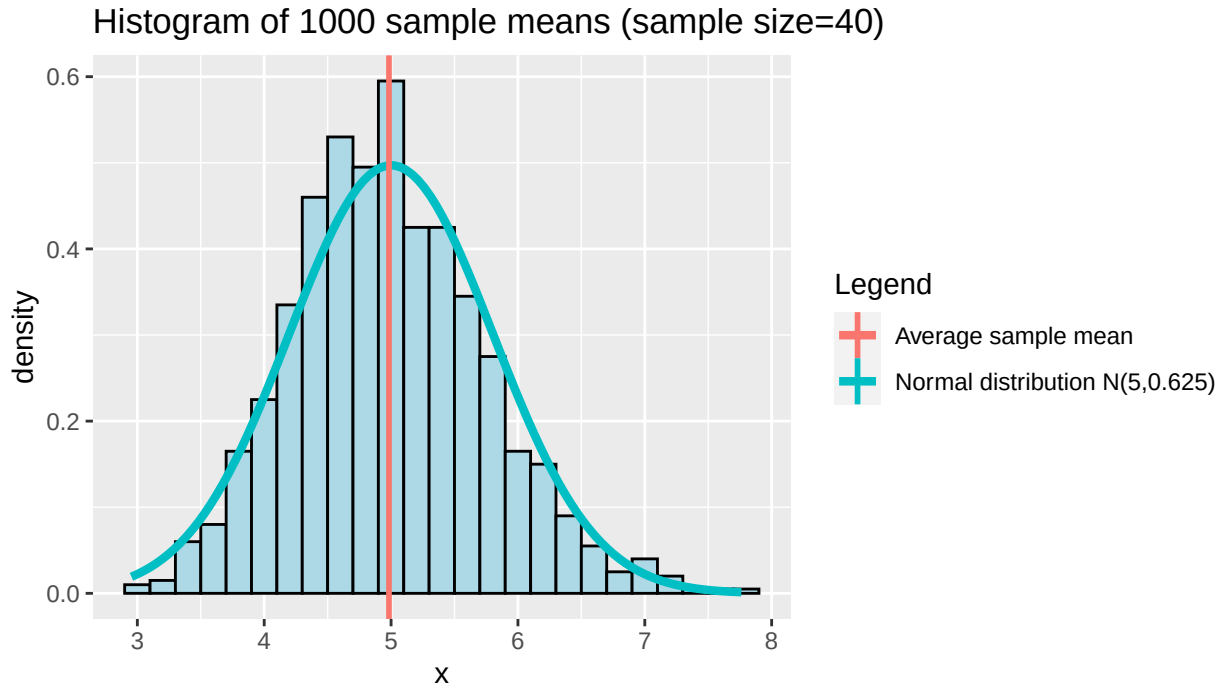

Figure 2: Plot 2

In conclusion, the distribution of means of 40 exponentials behave as predicted by the CTL

NOTE: The CTL doesn't tell us anything about the number of repeated samples to use, but only about the sample size. We should nonetheless be aware of the fact that *"when a simulation is run more than once, different results are obtained. Here we call this between-simulation variability Monte Carlo error (MCE)"* (from On the Assessment of Monte Carlo Error in Simulation-Based Statistical Analyses). This kind of error depends on the number B of random samples generated, tending to zero as as B tends to ∞. If for example we take 1000 new random samples, different from the ones we used before, their mean will be slightly different:

```
mns2<-NULL; for (i in 1 : 1000) mns2 = c(mns, mean(rexp(40,.2))); mean(mns2)- mn2
```

```
## [1] -0.001015016
```

With 10000 simulated samples instead of 1000 this difference would become even smaller.

Appendix

R Code for Table 1

This code is intended to be used in R Markdown with the chunck option "results='asis' ".

```
set.seed(241016)
x=rexp(1000,.2)
mu<-1/.2
sigma2<-1/.2^2
m<-log(2)/.2
mn=mean(x)
mdn<-median(x)
variance<-var(x)

t<-data.frame("Population Value"=c(
    paste("$\\mu$=",mu),paste("$\\sigma^2$=",sigma2),
    paste("$m$=",round(m,2))), "Sample Value"=c(paste("$\\bar{X}$=",round(mn,2)),
                                                paste("$s^2$=",round(variance,2)),
                                                paste("$m[X]$=",round(mdn,2))))
row.names(t)<-c("Mean", "Variance", "Median")
knitr::kable(t,
             caption="Population parameters and sample statistics for a random  sample of size=1000")
```

R Code for Plot 1

```
library(ggplot2)
f<-function(x) dexp(x,.2)
line.data <- data.frame(xintercept = c(mdn, mn), Lines = c("Sample median", " Sample mean"))
g<-ggplot(data.frame(x=x,Legend="Exponential pdf f(x;0.2)"), aes(x, col=Legend))
g<-g+geom_histogram(aes(y=after_stat(density)),col="black", fill="lightblue", binwidth = .5)
g<-g+stat_function(fun=f, lwd=1.5)
g<-g+geom_vline(aes(xintercept = xintercept, color = Lines), line.data, size = 1)
g<-g + ggtitle("Exponential distribution with rate=0.2")
g
```

R Code for Table 2

Once again, it works with results='asis' in R Markdown:

```
mns = NULL
for (i in 1 : 1000) mns = c(mns, mean(rexp(40,.2)))
mn2=mean(mns)
mdn2<-median(mns)
variance2<-var(mns)
t<-data.frame("Population Value"=c(
    paste("$\\mu$=",mu), paste("$\\sigma^2$=",sigma2/40),
    paste("$m$=",mu)), "Sample Value"=c(paste("$\\bar{X}$=",round(mn2,2)),
                                        paste("$s^2$=",round(variance2,2)),
                                        paste("$m[\\bar{X}]$=",round(mdn2,2))))
row.names(t)<-c("Mean", "Variance", "Median")
knitr::kable(t, caption="Population parameters and $\\bar{X}$ statistics for 1000 sample of size 40")
```

R Code for Plot 2

```
f<-function(x) dnorm(x,mu, sqrt(variance)/sqrt(40))
line.data <- data.frame(xintercept =mn2, Line = "Average sample mean")
g<-ggplot(data.frame(x=mns,Legend="Normal distribution N(5,0.625)"), aes(x, col=Legend))
g<-g+geom_histogram(aes(y=after_stat(density)),col="black", fill="lightblue", binwidth = .2)
g<-g+stat_function(fun=f, lwd=1.5)
g<-g+geom_vline(aes(xintercept = xintercept, color=Line), line.data, size = 1)
g<-g+ggtitle("Histogram of 1000 sample means (sample size=40)")
g
```