Investigating the Exponential Distribution

Jacques wagstaff

17 November 2016

Overview

The exponential distribution $f(x) = \lambda e^{-\lambda x}$, for $x \geq 0$, is characterised by one parameter: the rate parameter λ. The theoretical mean of the distribution is given by $\mu = 1/\lambda$ and the standard deviation is also $\sigma = 1/\lambda$.

The Central Limit Theorem (CLT) states that

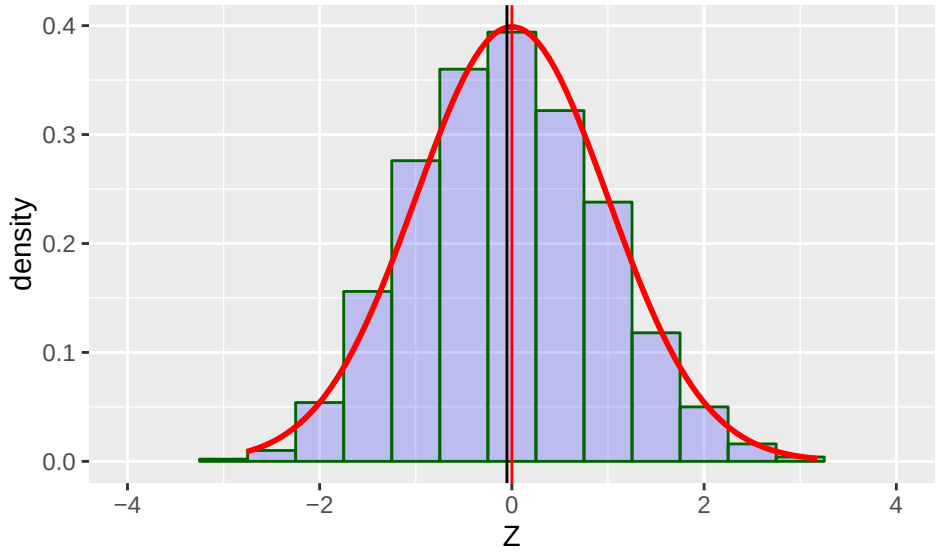
$$Z_n \equiv \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \lambda\sqrt{n}\left(\bar{X}_n - 1/\lambda\right)$$

has a distribution of a standard normal $N\left(\mu = 0, \sigma = 1\right)$, where $\bar{X}_n$ is the sample average, if n is large enough.

We investigate the distribution of averages of 40 exponentials, $n = 40$, and run a thousand simulations, $nosim = 1000$, to find the distribution of Z_n. We set $\lambda = 0.2$ for all of the simulations. The R code for the simulation is given below.

In the code below, we sample $n \times nosim$ numbers from the exponential distribution to simulate our data. We then take the average in groups of $n = 40$. The product $Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ is calculated for each group and the results are then plotted.

```
nosim<-1000; n<-40; lambda<-0.2; library(ggplot2); set.seed(1286)
cfunc <- function(x, n) sqrt(n)*(mean(x)-1/lambda)/(1/lambda)
# next we sample from the exponential distribution to simulate our data
dat <- data.frame(Z = apply(matrix(rexp(n*nosim,lambda),nosim), 1, cfunc, n))
g <- ggplot(dat, aes(x = Z)) + geom_histogram(alpha = .20, binwidth=.5,
                      colour = "darkgreen", fill = "blue", aes(y = ..density..))
g <- g + stat_function(fun = dnorm, size = 1,colour="red")
g<-g +geom_vline(xintercept = 0, colour="red") # The theoretical mean 1/lambda
g<-g +geom_vline(xintercept = mean(dat$Z)) # The sample mean
g<-g+coord_cartesian(xlim = c(-4,4))
g
```



The plot above shows a histogram for the density of $Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$. Overlaying the histogram we have also plotted the density function of a standard normal (line in red). The plot also shows two vertical lines showing the theoretical mean (in red) and the sample mean (in black).

1. The sample mean and the theoretical mean

For the sample mean and its interval we can perform a t.test:

```
t.test(dat$Z)
```

```
##
##  One Sample t-test
##
## data:  dat$Z
## t = -1.6595, df = 999, p-value = 0.09733
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.111637297  0.009334749
## sample estimates:
##   mean of x
## -0.05115127
```

The sample mean, also shown in the plot above as a vertical line in black, is given by -0.051 which is very close to the expected theoretical mean of a standard normal distribution i.e. 0. The 95% confidence interval for the mean is $(-0.112, 0.009)$, which includes 0. Note that the t-test assumes iid Gaussian variables.

2. The sample variance and the theoretical variance

The sample variance is given by

```
var(dat$Z)
```

```
## [1] 0.9500816
```

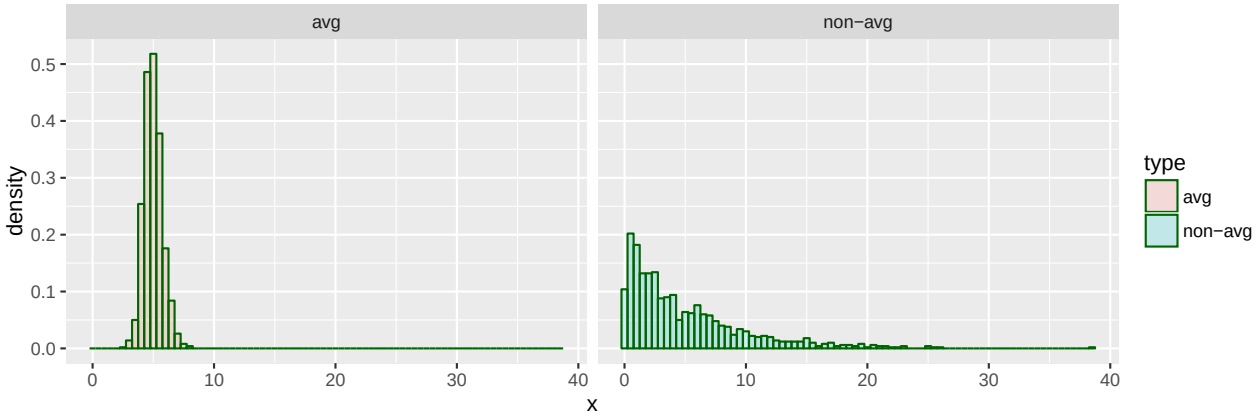
which is very close to the expected theoretical variance of a standard normal distribution i.e. 1.

3. Is the distribution Normal?

The first plot shown above, which shows a histogram for the density of Z_n and an overlay of the density function of a standard normal (line in red), clearly shows that the distribution is to a good approximation a standard normal. This is also confirmed by the previous calculations of the mean and variance of the sample which are to a good approximation those of a standard normal.

To illustrate this point further, we show in the plot below the difference between the distribution of a large collection of random exponentials (non-avg) and the distribution of a large collection of averages of 40 exponentials (avg).

```
dat1 <- data.frame(x = c(apply(matrix(rexp(n*nosim,lambda),nosim), 1, mean),
        rexp(nosim,lambda)),
        type = factor(rep(c("avg", "non-avg"), rep(nosim, 2))))
g1 <- ggplot(dat1, aes(x = x, fill = type)) +
        geom_histogram(alpha = .20, binwidth=.5, colour = "darkgreen",
                       aes(y = ..density..))
g1 + facet_grid(. ~ type)
```



Clearly the underlying distribution of the simulated data (non-avg) is not Gaussian (we know it comes from an exponential distribution), however the distribution of averages (avg) is clearly much more Gaussian.

Conclusion

We conclude that: **The distribution of means of 40 exponentials behave as predicted by the Central Limit Theorem**.