Course Project, Statistical Inference: Question 1

Felipe Campelo, Ph.D.

December 19, 2014

Abstract

This document presents the solution for Course Project Question 1, of the Statistical Inference course on Coursera.

Problem Statement

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 also *1/lambda*. Set ***lambda = 0.2*** for all of the simulations. In this simulation, you will investigate the distribution of averages of 40 exponential(0.2)s. Note that you will need to do a thousand or so simulated averages of 40 exponentials.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponential(0.2)s. You should:
1. Show where the distribution is centered at and compare it to the theoretical center of the distribution.
2. Show how variable it is and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.

Note that for point 3, focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

Solution

I'll address the three points defined in the previous section sequentially. First, I'll define the fixed parameters that will be used for this simulation exercise:

```
# Set the seed for the PRNG (for reproducibility)
set.seed(12345)

# Define lambda parameter for the exponential distributions.
lambda<-0.2

# Define the sample size to be used to obtain each estimate of the mean
n.obs<-40

# Define number of simulations to be performed:
n.sim = 9999
```

We can now proceed to address the points of this question:

> **Show where the distribution is centered at and compare it to the theoretical center of the distribution.**

According to the Central Limit Theorem (CLT), the distribution of the sample mean will have the following three properties (assuming a large enough and independent sample. Both assumptions are verified in this case):

(i) it will be (approximately) Gaussian; (ii) its mean will be equal to the mean of the original distribution, i.e., equal to *1/lambda = 1/0.2 = 5*; and (iii) its variance will be equal to the variance of the original distribution divided by the sample size, i.e. $(1/lambda)^2/n.obs = 1/(0.04 \times 40) = 0.625$.

I demonstrate how closely the empirical values come to the theoretical ones below:

```
# Simulate the data:
simexp<- matrix(
      data = rexp(n=n.sim*n.obs,
                  rate=lambda),
      ncol = n.sim)

# Get the sample mean from each set of 40 independent observations
simmean<-colMeans(simexp)

# Compare the center of the empirical distribution with the theoretical value.
# Remember that the theoretical value in this case is 1/lambda = 1/0.2 = 5
mean(simmean)
```

```
## [1] 4.99707
```

As we can see, it is pretty close to the reference (theoretical) value.

Show how variable it is and compare it to the theoretical variance of the distribution

This can also be very easily shown:

```
# Compare the variability of the empirical distribution with the theoretical value.
# Remember that the theoretical value in this case is (1/lambda^2)/n.obs = 0.625
var(simmean)
```

```
## [1] 0.6104097
```

Again, a result that is quite close to the reference value.

Show that the distribution is approximately normal.

I will address this question using a qualitative approach, namely a graphical comparison between the original distribution (*Exp(lambda = 0.2)*) against the sample distribution of means of 40 observations from *Exp(lambda = 0.2)*, against the theoretical distribution predicted by the CLT (*Normal(mean=5,var=0.625)*). The code and resulting figure are shown below:

```
# Set graphical parameters
par(xpd=FALSE, oma=c(1.5,0,0,2.5),mfcol=c(2,2),
    mar=c(5,4,4,3)+0.1, mgp=c(1.8,.2,0),
    tck=0.02, bg="#DDF0F0")

library(beanplot)

# Fig 1-a: Density estimation for Exp(0.2)
beanplot(rexp(n=9999,rate=lambda),log="",what=c(F,T,T,F),
         col="lemonchiffon",side="second",horizontal=T,
```

```
         main="Exp(0.2)",xlab="x",ylab="Density(x)")

# Fig 1-b: Density estimation for mean(40 Exp(0.2))
beanplot(simmean,log="",what=c(F,T,T,F),
         col="lemonchiffon",side="second",horizontal=T,
         main="mean of 40 Exp(0.2)",xlab="x",ylab="Density(x)")

# Fig 1-c: Normal qq plot for mean(40 Exp(0.2))
qqnorm(simmean,pch=16,main="Normal QQ plot for mean of 40 Exp(0.2)")
qqline(simmean,col="red")

# Fig 1-d: Histogram and True Gaussian
h<-hist(simmean,col="darkgray",border="white",breaks=25,
        main="Sample x Theoretical distributions",xlab="x")
xfit<-seq(min(simmean),max(simmean),length=40)
yfit<-dnorm(xfit,mean=5,sd=sqrt(0.625))
yfit <- yfit*diff(h$mids[1:2])*length(simmean)
lines(xfit, yfit, col="red", lwd=3)
```

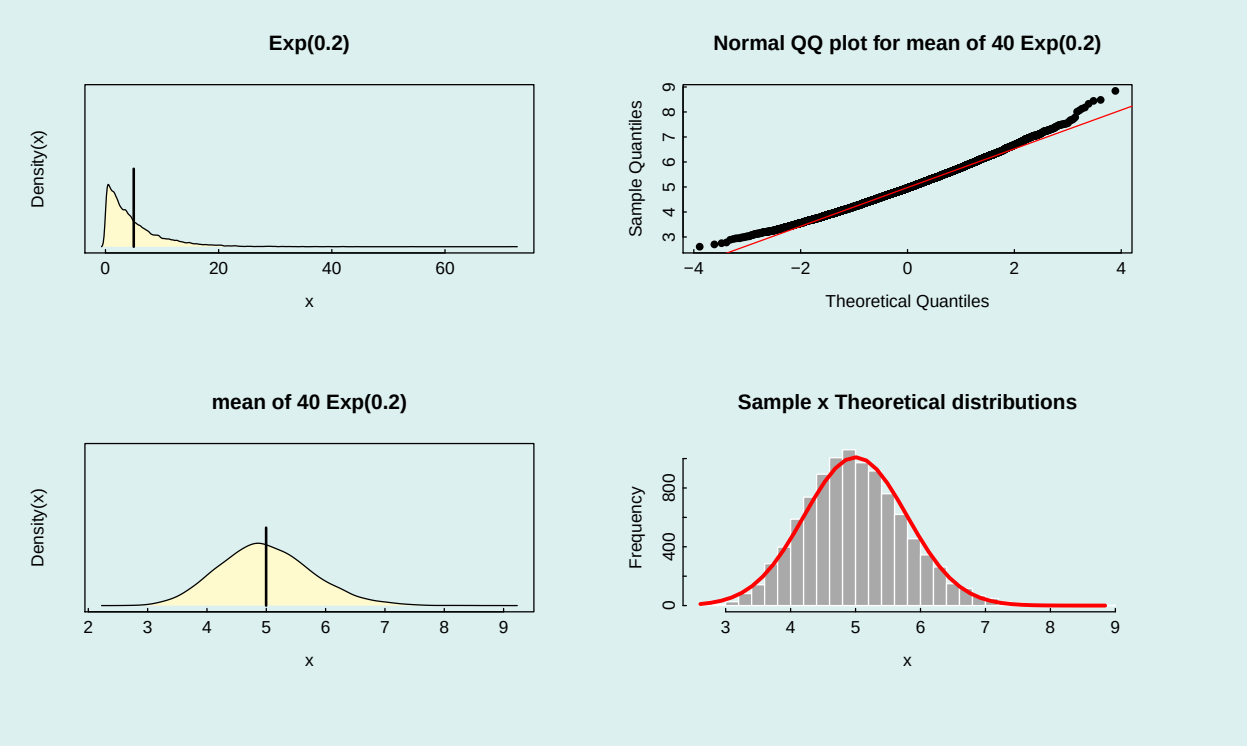

Figure 1:These four panels show very clearly that, despite the very skewed distribution of *Exp(0.2)*, the resulting sampling distribution of the means of 40 observations of this variable approximates a normal distribution with mean and variance closely matching the theoretical predictions.