DSDM2015

Coursera Data Science Specialization, Statistical Inference Class (November 2015)

Assignment 1: The Exponential Distribution and the Central Limit Theorem

According to the Central Limit Theorem, the sampling distribution of the mean of a number of random samples will follow a normal distribution if the sample size is large enough. The mean of the sampling distribution will converge to the population mean and the variance will converge to the population variance divided by the sample size as the sample size increases to infinity. We will demonstrate the effect of the central limit theorem using simulations. The population will follow an exponential distribution (with known mean and variance). The exponential distribution is quite different from the normal distribution, which provides us with a strong demonstration of the effect of the central limit theorem. Thus, the following will show a) that the sampling distribution follows a normal distribution (question 3) and b) that the mean and variance of the sampling distribution in the simulation are quite close to the theoretical values (question 1 and 2).

The exponential distribution is often used to model the duration of a process. It has the property of assigning a positive density to all nonnegative numbers. There is a large family of exponential distributions; a specific exponential distribution is determined by its rate (lambda). Figure 1 shows the exponential distribution (blue line) with lambda set at 0.2. The mean of the exponential distribution is given by 1/lambda and its variance is 1/lambda^2 (hence the standard deviation would be 1/lambda, like the mean). Both the standard deviation and the mean (dotted blue line) are equal to five in our example.

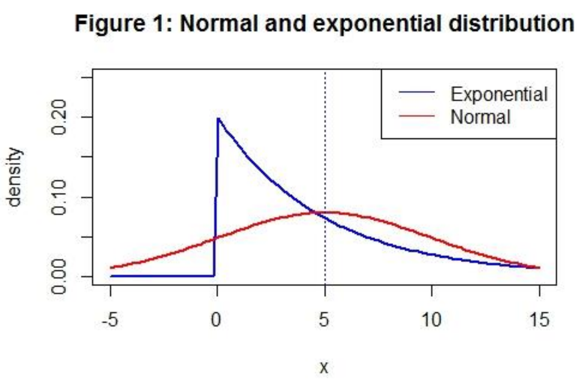
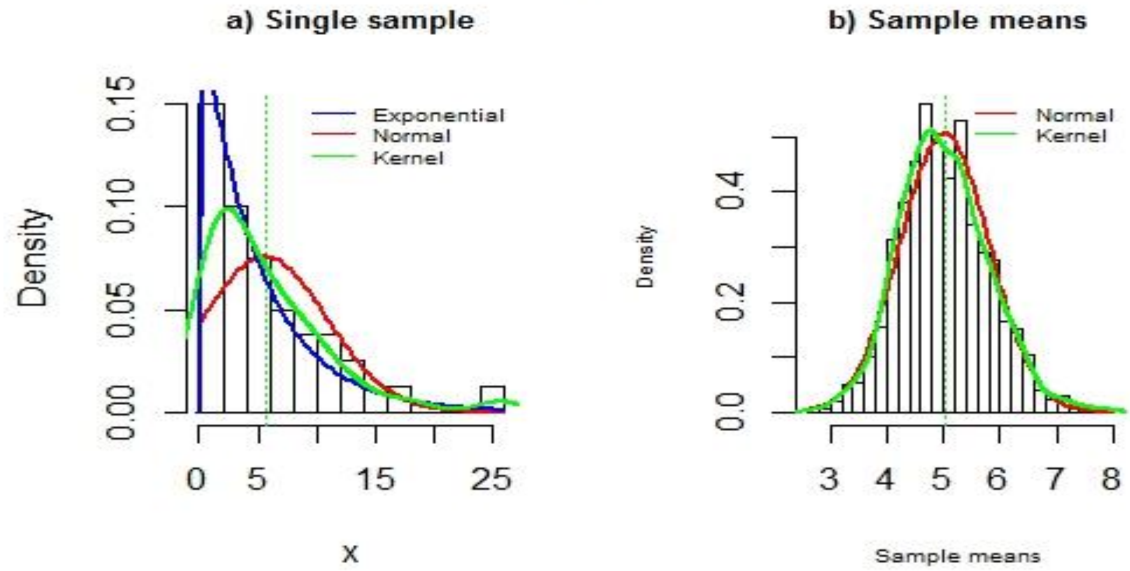


Figure 1 also shows the normal distribution (with the same mean and standard deviation) (red line). Comparing the two distributions we note that they are quite different. The normal distribution is symmetrical around the mean and also puts positive density on negative values whereas the exponential function is highly asymmetrical.

To study the behavior of the sampling distribution of the exponential distribution, we will draw 1000 random samples of size 40 from the exponential distribution described above. Figure 2, panel a shows one of these samples. Superimposed on the figure we can see a kernel density estimate for the sample (green line) as well as an exponential (with lambda=0.2) and a fitted normal distribution. As we can see, drawing a single sample from an exponential distribution results in a set of values that roughly follows the distribution of the population from which it was drawn. In other words, it looks like the exponential distribution (blue line). The red line shows the normal distribution with the mean and standard deviation taken from this sample. This would be a very poor approximation of the sample distribution.

Figure 2



Panel b shows the results of 1000 simulations (see Appendix for the R code). The histogram shows the distribution of the average of the sample means. Superimposed are the normal distribution (using the empirical mean and standard deviation) (red line) and a kernel density estimate (green line). The distribution of the mean of the sample means, i.e., the sampling distribution of the mean, is well approximated by a normal distribution.

Figure 3

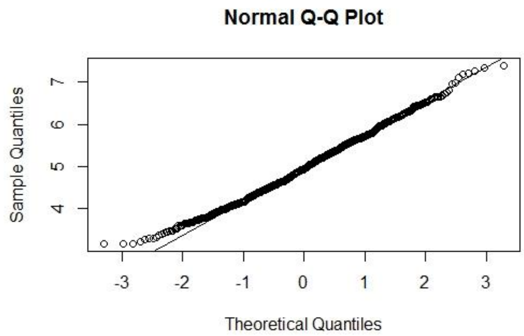


Figure 3 shows a plot comparing the quantiles of the sampling distribution to the normal distribution. As one can see, the sampling distribution from the simulation closely follows a normal distribution.

As noted above, the theoretical mean of an exponential distribution is 1/lambda. In our example (lambda=0.2), this value is five. According to the central limit theorem, the sampling distribution should have the same value. The simulation has indeed a mean of sample averages that is close to 5 (Table 5).

Table 1: Comparing theoretical and simulated mean and variance of the sampling distribution		
	Theoretical	Simulation (estimate)
Mean (sampling distribution)	5	4.974
Variance (population)	25	22.826
Variance (sampling distribution)	0.625	0.571

The variance of the exponential distribution is $1/lambda^2$. In our example, this would be 25 for the population from which the samples are drawn. To derive the variance of the sampling distribution, we have to divide the population variance by the sample size. In our case, this would result in a theoretical variance of 0.625 for the sampling distribution for lambda 0.2 and sample size 40. The simulation results are quite close to these values (see Table 1). The estimated population variance is 22.826 and the sampling distribution in the simulations is .571.

Thus, the simulation shows that the sampling distribution of the exponential distribution is normal and its mean and variance are close to the theoretical values. Running more simulations would reduce the difference between the values even further. In addition, the estimates from the simulations would be closer to the actual (theoretical) values if the sample size would be larger.

Appendix (R Code)

```
# Simulation of Exponential Random Varables and CLT
#
# Comparing exponentional and normal distribution

set.seed(1234)

exp.mean<-1/0.2
x<-seq(-5,15, length=100)
y<-seq(0,0.25, length=100)
y.exp<-dexp(x, rate=0.2)
y.norm<-dnorm(x, mean=exp.mean, sd=exp.mean)
plot(x,y,type="n",lwd=2,ylab="density", main="Figure 1: Normal and exponential distribution")
lines(x,y.exp,type="l",lwd=2,col="blue")
lines(x,y.norm,type="l",lwd=2,col="red")
abline(v=exp.mean,col="blue",lty=3)
legend("topright", c("Exponential", "Normal"), lty=c(1, 1), col=c("blue", "red"))

# Simulation results
nsim<-1000
nsamples<-40
lambda<-0.2
my.means<-numeric(nsim)

for (i in 1: nsim) {
  new.sample<-rexp(nsamples, lambda)
  my.means[i]<-mean(new.sample)
}

# Graph comparing a single sample to simulation results
par(mfrow = c(1, 2))
single.sample<-rexp(nsamples, lambda)
hist(single.sample, breaks=10, main="a) Single sample", cex.main=0.8, xlab="x", prob=T)
mtext('Figure 2', outer=T, line=-1, cex=1.4)
abline(v=mean(single.sample),col="green",lty=3)
legend("topright", cex=0.6, bty="n", c("Exponential", "Normal", "Kernel"), lty=c(1, 1, 1), col=c("blue",
"red", "green"))
#Fitting normal distribution
x<-seq(min(single.sample),max(single.sample),length=40)
curve(dnorm(x, mean=mean(single.sample), sd=sd(single.sample)), col="red", lwd=2, add=TRUE)
# Fitting exponential distribution
curve(dexp(x, rate=lambda), col="blue", lwd=2, add=TRUE)
#Fitting kernel density
lines(density(single.sample), col="green", lwd=2)
hist(my.means, breaks=20, main="b) Sample means", cex.main=0.8, cex.lab=0.6, cex.lab=0.6,
xlab="Sample means", prob=T)
abline(v=mean(my.means),col="green",lty=3)
x<-seq(min(my.means),max(my.means),length=40)
curve(dnorm(x, mean=mean(my.means), sd=sd(my.means)), col="red", lwd=2, add=TRUE)
lines(density(my.means), col="green", lwd=2)
legend("topright", cex=0.6, bty="n", c("Normal", "Kernel"), lty=c(1, 1), col=c("red", "green"))
```

```
#Showing normality (q-q-plot)
dev.off()
qqnorm(my.means)
qqline(my.means)
mtext('Figure 3', outer=T, line=-1, cex=1.4)

# Comparaing theoretical and simulation results
1/0.2
mean(my.means)
1/0.2^2
var(my.means)*nsamples
```