Statistical Inference Project

JP Dunlap

July 27, 2017

Part 1: Simulation Exercise

Investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution is 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 1/lambda. Set lambda = 0.2 for all of the simulations. Investigate the distribution of averages of 40 exponentials. Conduct 1000 simulations.

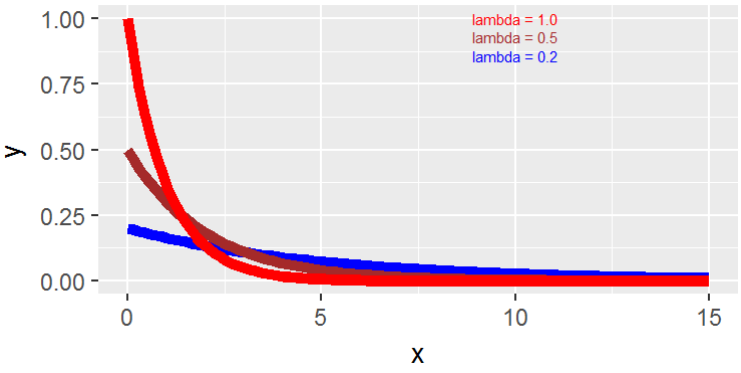
```
## Initialize lambda, mu, and sigma

lambda <- 0.20
mu <- 1/lambda
sigma <- 1/lambda
```

Discussion

The exponential distribution is the probability distribution in which events occur continuously and independently based on a constant rate of change. It describes the time between events in a Poisson process (https://en.wikipedia.org/wiki/Exponential_distribution (https://en.wikipedia.org/wiki/Exponential_distribution)).

The exponential distribution has a theoretical mean and standard deviation equal to 1/lambda, where lambda is the rate of change in the underlying Poisson process. This means that the shape of the exponential distribution is completely dependent on the rate of change (lambda) as illustrated in this plot.

```
## use ggplot to create replication of three exponential distributions with three different lambdas.

g <- ggplot(data.frame(x=c(0,15)),aes(x=x)) + theme_grey()
g <- g + stat_function(fun=dexp,geom = "line",size=2,col="blue",args = (rate=.2))
g <- g + stat_function(fun=dexp,geom = "line",size=2,col="brown",args = (rate=0.5))
g <- g + stat_function(fun=dexp,geom = "line",size=2,col="red",args = (rate=1.0))
g <- g + annotate("text",x=10,y=1.0,label = "lambda = 1.0", color = "red", cex = 2)
g <- g + annotate("text",x=10,y=.93,label = "lambda = 0.5", color = "brown", cex = 2)
g <- g + annotate("text",x=10,y=.86,label = "lambda = 0.2", color = "blue", cex = 2)
g
```



The impact of the changes in lambda are clear from the above chart.

Simulation

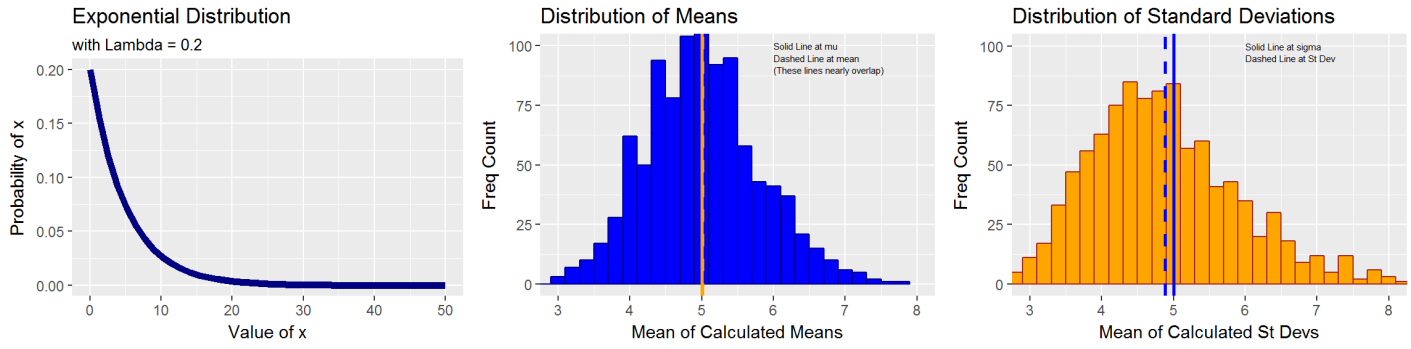
The task at hand is to simulate 1000 exponential distributions with 40 observations each to determine of the Central Limit Theorem is support for exponential distributions. That is, test the distribution of the means and standard deviations of 1000 simulations of of 40 values from taken from an exponential distribution with Lambda = 0.20, to determine if they approximate the theoretical mu and sigma of the distribution, 1/lambda = 1/0.20 = 5.

```
## perform calculations for simulation

## xx is a sequence of 10000 values between 0 and 50
xx <- seq(0, 50, length.out=40)

## expDat is a data frame containing x and y coordinates of an exponential distribution, lambda = 0.2
expDat <- data.frame(xx=xx, yy=dexp(xx, rate=0.2))

## simsd is the standard deviation of each simulation
## simmn is the mean of each simulation

simmn <- rep(NULL, time=1000)
simsd <- rep(NULL, time=1000)

## run 1000 simulations of 40 values each
for (i in 1 : 1000){
        simmn[i] <- mean(rexp(n = 40, rate = 0.20))
        simsd[i] <- sd(rexp(n = 40, rate = 0.20))
}

mean.simmn <- mean(simmn)
mean.simsd <- mean(simsd)
```

If the CLT holds, one would expect to see the mean of the simulated means and the mean of the simulated standard deviations to be approximately equal to the theoretical value of 5 (1/lambda = 1/.2 = 5.0). The actual calculated mean of the simulated means is 5.0108542, and the calculated mean of the simulated standard deviations is 4.8833921.

The following three graphs show: 1) the actual distribution of a sample of 40 random observations taken from the exponential probability distribution, 2) a histogram of the mean of the simulated means, 3) and the mean of the simulated standard deviations. Both 2) and 3) are taken from 1000 samples of 40 random observations taken from the same exponential probablility distribution as 1).

```
f <- ggplot(expDat, aes(x=xx, y=yy)) + geom_line(color = "navy", lwd = 2) + theme_grey()
f <- f + labs(x = "Value of x", y = "Probability of x", title = "Exponential Distribution", subtitle = "with Lambda = 0.2")

g <- ggplot(data.frame(simmn), aes(simmn)) + geom_histogram(binwidth = 0.20, col = "navy", fill = "blue")
g <- g + labs(x = "Mean of Calculated Means", title = "Distribution of Means", y = "Freq Count")
g <- g + coord_cartesian(xlim=c(3,8), ylim=c(0,100)) + theme_grey()
g <- g + geom_vline(xintercept = 5, col = "orange", lwd = 1)
g <- g + geom_vline(xintercept = mean(simmn), col = "orange", lwd = 1, lty = 2)
g <- g + annotate("text",x=6,y=100,label = "Solid Line at mu", color = "black", cex = 2,hjust=0)
g <- g + annotate("text",x=6,y=95,label = "Dashed Line at mean", color = "black", cex = 2,hjust=0)
g <- g + annotate("text",x=6,y=90,label = "(These lines nearly overlap)", color = "black", cex = 2,hjust=0)
h <- ggplot(data.frame(simsd), aes(simsd)) + geom_histogram(binwidth = 0.20, col = "brown", fill = "orange")
h <- h + labs(x = "Mean of Calculated St Devs", title = "Distribution of Standard Deviations", y = "Freq Count")
h <- h + coord_cartesian(xlim=c(3,8), ylim=c(0,100)) + theme_grey()
h <- h + geom_vline(xintercept = 5, col = "blue", lwd = 1)
h <- h + geom_vline(xintercept = mean(simsd), col = "blue", lwd = 1, lty = 2)
h <- h + annotate("text",x=6,y=100,label = "Solid Line at sigma", color = "black", cex = 2,hjust=0)
h <- h + annotate("text",x=6,y=95,label = "Dashed Line at St Dev", color = "black", cex = 2,hjust=0)
plot_grid(f,g,h,nrow = 1, ncol = 3)
```



Part 1 Conclusion

The first graph in this sequence displays the sample distribution of 40 observations on an exponential distribution with Lambda = 0.20. This distribution does not appear Gaussian at all.

However, both the calculated statistics and the two histograms show what appears to be a far more Gaussian distribution of the mean and of the standard deviation about the theoretical values of 5 for both mu and sigma. It appears reasonable to conclude that the Central Limit Theorem holds in the case of exponential distributions.

The actual calculated mean of the simulated means is 5.0108542 compared to a theoretical value of 5.0. The calculated mean of the simulated standard deviations is 4.8833921 compared to a theoretical value of 5.0.