## Introduction to the Central Limit Theorem (CLT)

This is my reproducible markdown script for peer assignment 1 part A of the Coursera course Statistical In ference.

By Robert Merriman

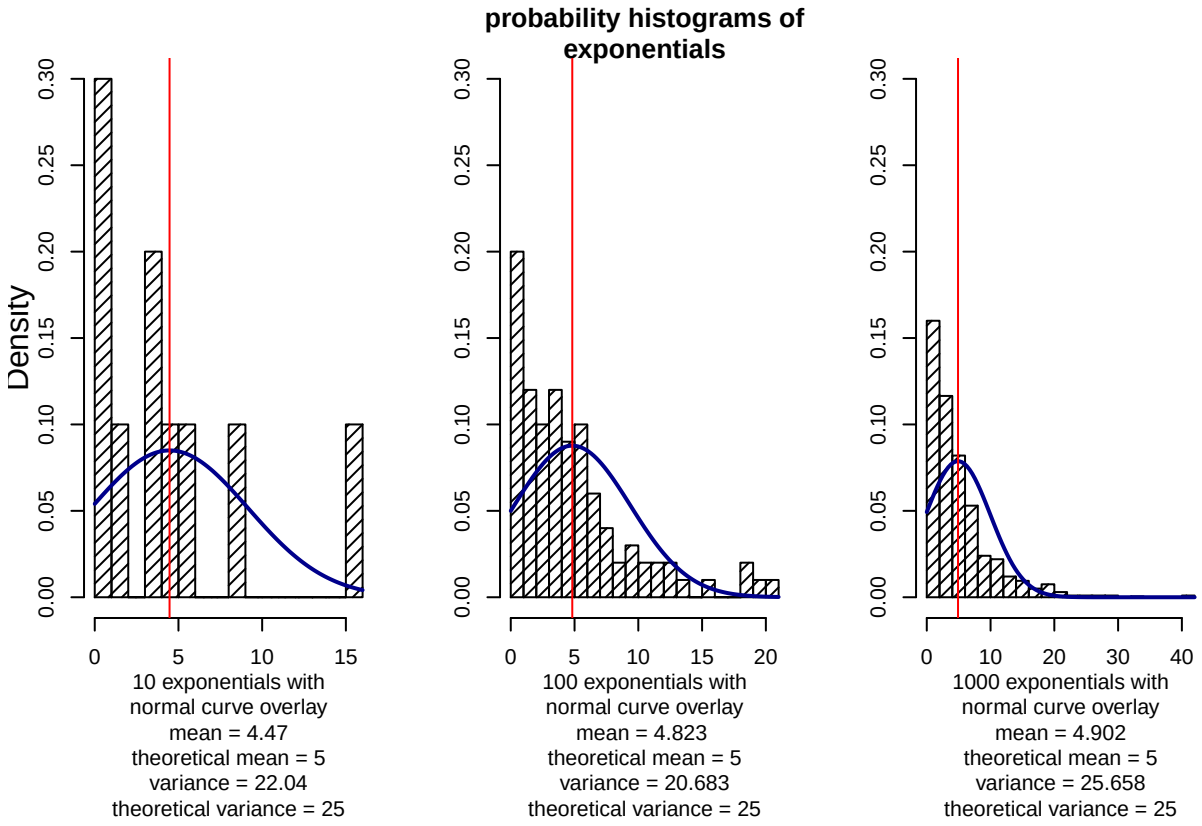
Course reference URL

In this project I will investigate the exponential distribution in R and compare it with the Central Limit Theorem Reference: "*In probability theory, the central limit theorem (CLT) states that, given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed, regardless of the underlying distribution.*"

The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. For this assignment lambda = 0.2 for all simulations. The theoretical mean of the exponential distribution is 1/lambda and the standard deviation is also 1/lambda. As lambda is .2, the theoretical mean and standard deviation are 1 / 0.2 = 5. Variance is the square of the standard deviation and for the exponential distribution with a lambda of 0.2 is 5^2 = 25.

The exponential distribution is not normally distributed as shown in the following density histograms where I've overlayed a normal distribution curve and indicated the mean of the exponential distributions with a red vertical line:

```
set.seed(100)
par(mfrow=c(1,3), mar=c(10,3,2,2), mgp=c(7,1,0))
sim1 <- data.frame(x = rexp(10, 0.2))
m1=mean(sim1$x)
sd1=sd(sim1$x)
sim2 <- data.frame(x = rexp(100, 0.2))
m2=mean(sim2$x)
sd2=sd(sim2$x)
sim3 <- data.frame(x = rexp(1000, 0.2))
m3=mean(sim3$x)
sd3=sd(sim3$x)
hist(sim1$x, density=20, breaks=20, prob=TRUE,
     xlab=paste0("10 exponentials with\nnormal curve overlay\nmean = ",
        round(m1,2),"\ntheoretical mean = 5", "\nvariance = ",
        round(sd1^2,3),"\ntheoretical variance = 25"), ylim=c(0, 0.3), main="")
mtext("Density",side=2,line=2)
curve(dnorm(x, mean=m1, sd=sd1), col="darkblue", lwd=2, add=TRUE, yaxt="n")
abline(v=m1,col="red")
hist(sim2$x, density=20, breaks=20, prob=TRUE,
      xlab=paste0("100 exponentials with\nnormal curve overlay\nmean = ",
        round(m2,3)," \ntheoretical mean = 5", "\nvariance = ",
        round(sd2^2,3),"\ntheoretical variance = 25"), ylim=c(0, 0.3),
        main="probability histograms of\nexponentials")
curve(dnorm(x, mean=m2, sd=sd2), col="darkblue", lwd=2, add=TRUE, yaxt="n")
abline(v=m2,col="red")
hist(sim3$x, density=20, breaks=20, prob=TRUE,
     xlab=paste0("1000 exponentials with\nnormal curve overlay\nmean = ",
        round(m3,3),"\ntheoretical mean = 5", "\nvariance = ",
        round(sd3^2,3),"\ntheoretical variance = 25"), ylim=c(0, 0.3), main="")
curve(dnorm(x, mean=m3, sd=sd3), col="darkblue", lwd=2, add=TRUE, yaxt="n")
abline(v=m3,col="red")
```



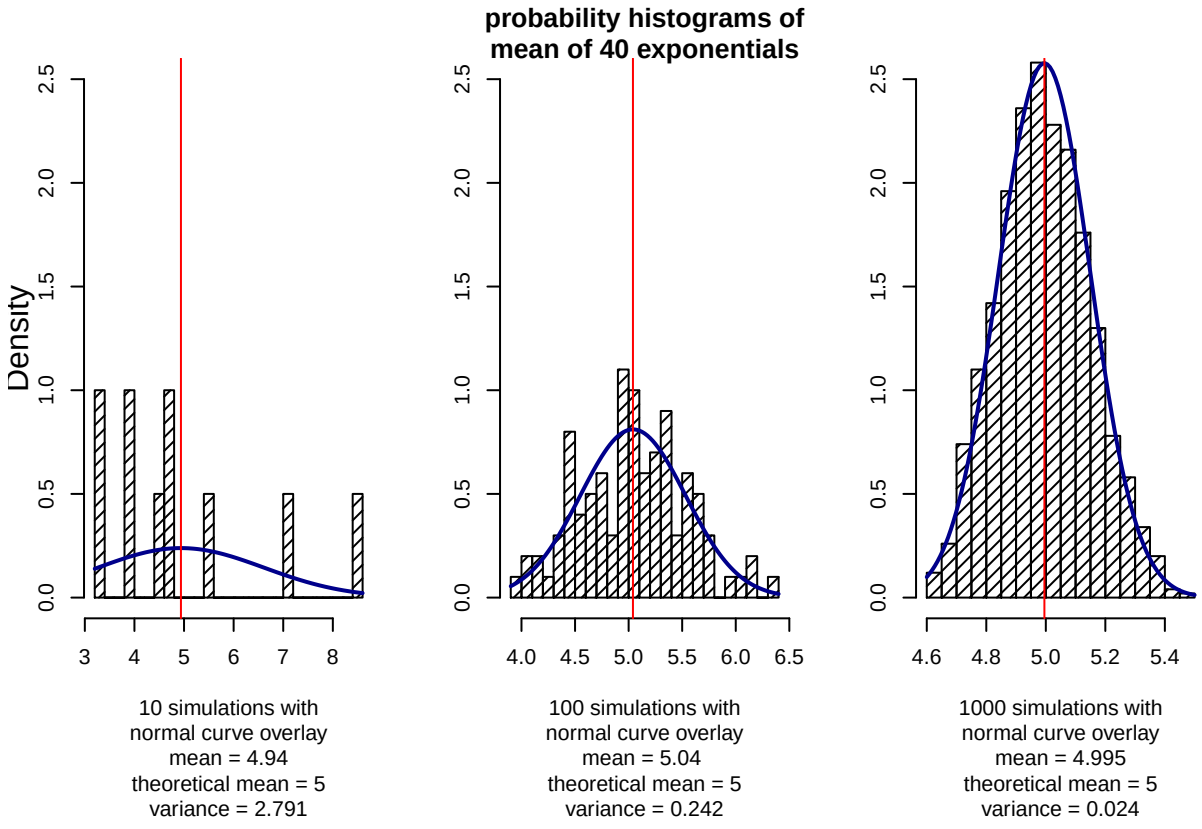

Below each histogram I have compared the mean, variance, theoretical mean, and the theoretical variance. As the number of exponentials increased the mean and the variance converged to the theoretical mean and theoretical variance.

According to the central limit theorem, the arithmetic mean of a large number of iterations of the exponential distribution should be approximately normally distributed. I will investigate the distribution of averages of 40 exponentials.

As the number of iterations increased the distribution did become approximately normally distributed as shown in the following histograms where I've overlayed a normal distribution curve and indicated the mean of the exponential distributions with a red vertical line:

```
set.seed(100)
par(mfrow=c(1,3), mar=c(10,3,2,2), mgp=c(7,1,0))
sim1 <- data.frame(x = sapply(1:10, function(x) {mean(rexp(10, 0.2))}))
m1=mean(sim1$x)
sd1=sd(sim1$x)
sim2 <- data.frame(x = sapply(1:100, function(x) {mean(rexp(100, 0.2))}))
m2=mean(sim2$x)
sd2=sd(sim2$x)
sim3 <- data.frame(x = sapply(1:1000, function(x) {mean(rexp(1000, 0.2))}))
m3=mean(sim3$x)
sd3=sd(sim3$x)
hist(sim1$x, density=20, breaks=20, prob=TRUE,
     xlab=paste0("10 simulations with\nnormal curve overlay\nmean = ",
        round(m1,2),"\ntheoretical mean = 5", "\nvariance = ",
        round(sd1^2,3)), ylim=c(0, 2.5), main="")
```

```
curve(dnorm(x, mean=m1, sd=sd1), col="darkblue", lwd=2, add=TRUE, yaxt="n")
abline(v=m1,col="red")
mtext("Density",side=2,line=2)
hist(sim2$x, density=20, breaks=20, prob=TRUE,
      xlab=paste0("100 simulations with\nnormal curve overlay\nmean = ",
        round(m2,3)," \ntheoretical mean = 5", "\nvariance = ",
        round(sd2^2,3)), ylim=c(0, 2.5),
        main="probability histograms of\nmean of 40 exponentials")
curve(dnorm(x, mean=m2, sd=sd2), col="darkblue", lwd=2, add=TRUE, yaxt="n")
abline(v=m2,col="red")
hist(sim3$x, density=20, breaks=20, prob=TRUE,
     xlab=paste0("1000 simulations with\nnormal curve overlay\nmean = ",
        round(m3,3),"\ntheoretical mean = 5", "\nvariance = ",
        round(sd3^2,3)), ylim=c(0, 2.5), main="")
curve(dnorm(x, mean=m3, sd=sd3), col="darkblue", lwd=2, add=TRUE, yaxt="n")
abline(v=m3,col="red")
```




Below each histogram I have compared the mean, theoretical mean, and the variance. As the number of simulations increased the mean of the averages of 40 exponentials approached the theoretical mean and was approximately normally distributed hence demonstrating the CLT. The variance of the 1000 iterations of the mean of 40 exponentials was significantly smaller than the variance of 1000 exponentials.