Justifying the Means

Investigation of the Central Limit Theorem Using Exponential Distribution

Coursera Student

Overview

The Central Limit Theorem (CLT) states that the means of a large number of samples of a distribution will tend toward a normal distribution even if the underlying observations are not normally distributed. We're going use an exponential distribution (not a normal distribution) to show an example of the CLT at work.

Simulations

An exponential distribution describes the time between events that occur continuously and independently at an average rate (a Poisson process). As an example, imagine a nearby hospital that averages 12 new babies/day. The exponential distribution is the time between births. It can be described with λ, where λ is the number of events in a given time frame. In our hospital example λ would be .5 babies/hour. With that definition, as can be easily imagined, the mean will be around $\frac{1}{\lambda}$ (again, using the hospital example, one baby every 2 hours). But it will not be normal; the time between events can never be negative, but it might stretch out indefinitely.

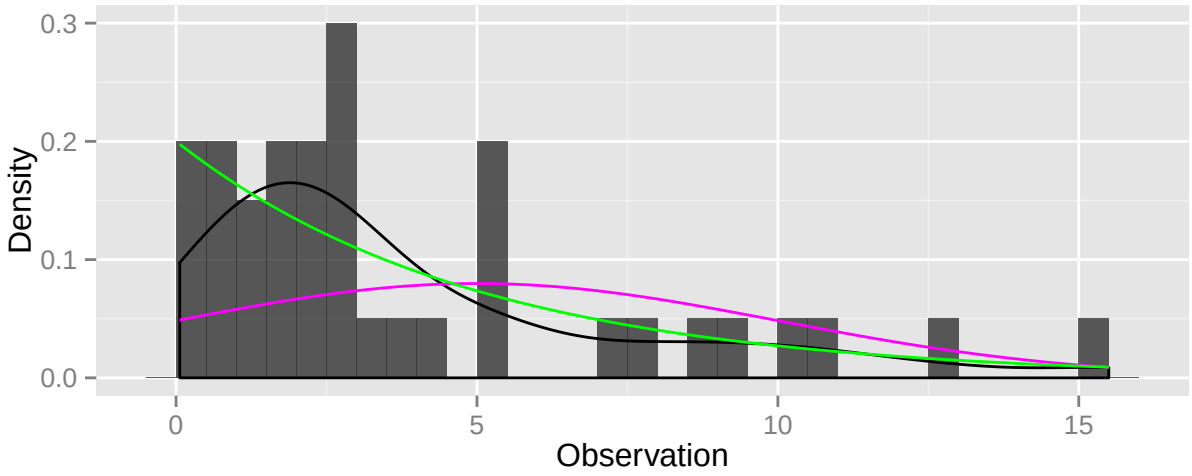
We're going to use R to create 1000 samples of 40 observations using an exponential distribution with λ=.2. We'll then use these samples to see if the CLT holds up for them. We begin by creating the samples:

```
set.seed(62015)
ourSimulations <-vector(length=1000,mode='list')
for(i in 1:1000){
  ourSimulations[[i]]<-rexp(40,.2)}
```

The numbers are mostly small, mostly clustered around 5 or so with a long tail, as expected based on the definition of this type of distribution.

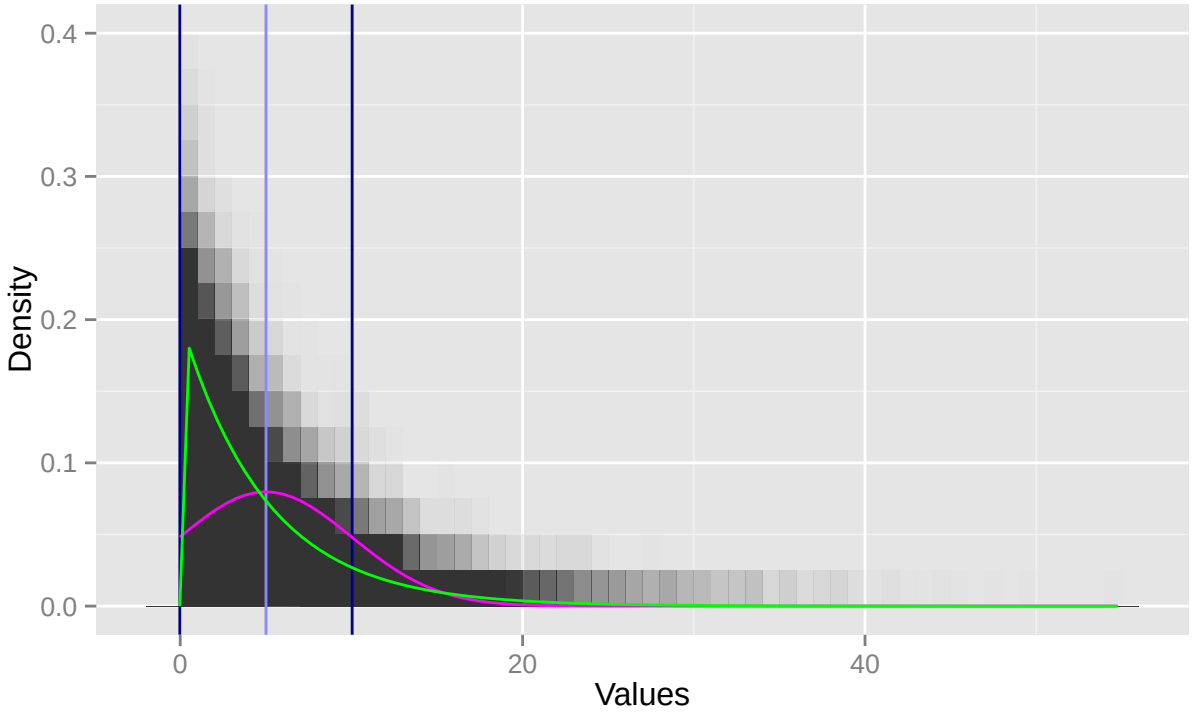
Samples, Theoreticals, Distributions

We can take a quick look at the histogram of one of these samples. We've superimposed the density in black, the normal function in purple, and the exponential function in green.



The plot makes it clear that it isn't a normal distribution, but it does look vaguely like the exponential distribution that we used to generate the sample.

For an even more thorough diagram, we can look at all of the histograms for all of the samples superimposed of each other. Again, we've graphed the normal curve in purple and the exponential curve in green.

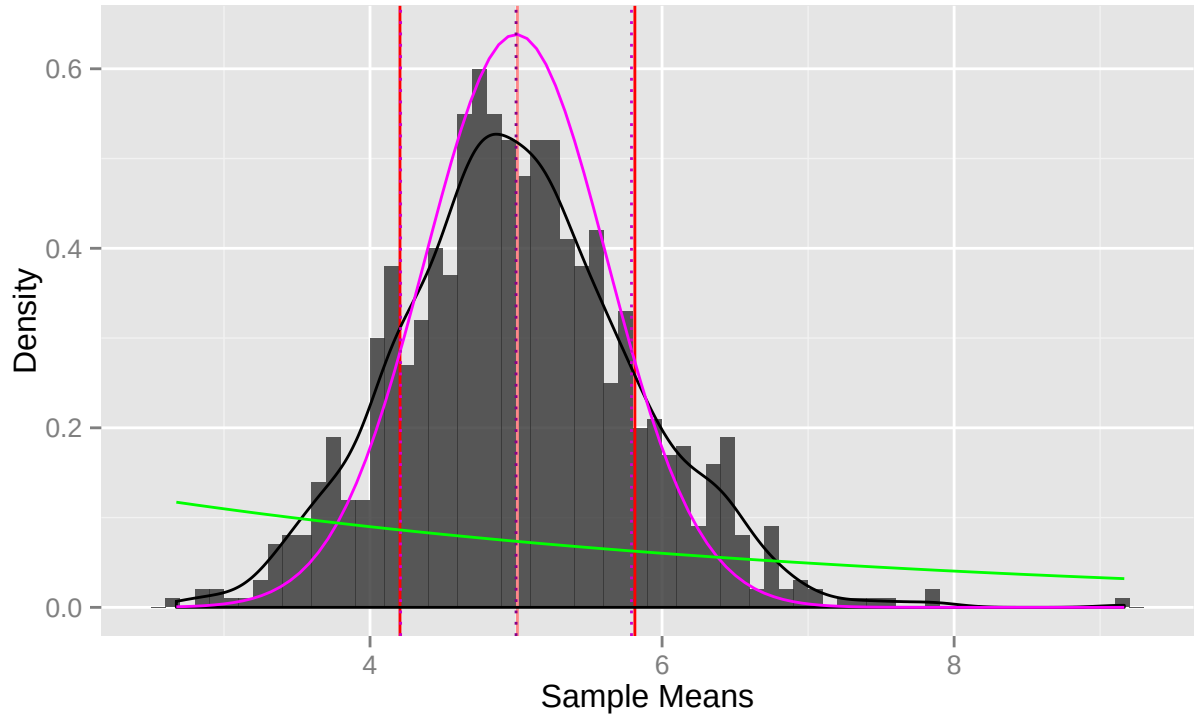


Altogether, it is clear that we're definitely not looking at a normal distribution, but our samples do track rather nicely to the exponential distribution curve. The mean is, as expected, right around 5, shown with the lighter blue line. The standard deviation is also around, shown with the darker blue lines. The variance is a little over 25.

Now, we take a look at the means of the experiments to see how they are distributed.

```
ourSimMeans <- vector(length=1000,mode='numeric')
for(i in 1:1000){
  ourSimMeans[i]<-mean(ourSimulations[[i]])}
```

We will graph the histogram of the means, again superimposing the normal curve in purple, the exponential curve in green, and the density in black. The mean is in light red with the standard deviations in dark red. The theoretical mean and standard deviations are in dotted lines.



As can be seen, the density is very close to the normal curve. There's no resemblance at all to the exponential curve. Clearly, the mean is still at the same point, right around 5. But the variance and standard deviation of the means is different than the variance and standard deviation of the samples. Variance is just about $\frac{\sigma^2}{40}$ (the population variance over the sample size). The standard deviation is the square root of that.
Here is the summary of our experiments

```
##                            Mean Variance Standard Deviation
## theoreticalStatistics      5.00   25.000             5.0000
## populationStatistics       5.01   25.330             5.0300
## theoreticalMeansStatistics 5.00    0.625             0.7906
## meansStatistics            5.01    0.650             0.8000
```

The theoretical mean, the mean of the population, and the mean of the means are all nearly the same (as expected by the CLT). The theoretical variance and variance of the population are similar, but the variance of the means is different. It is close to the population variance over the sample size.

Appendix

Packages

To complete this investigation, we use the following packages:

- reshape
- ggplot2
- plyr
- dplyr

Single Sample

Here, we pick a single sample and plot the histogram:

```
singleSample <- sample(1:1000,1)
ggplot()+aes(ourSimulations[[singleSample]])+
  geom_histogram(aes(y=..density..), bin=.5, alpha=.8)+
  geom_density(col="black")+
  xlab("Observation")+
  ylab("Density")+
  stat_function(fun=dnorm,args=list(mean=5, sd=5), color="#FF00FF")+
  stat_function(fun=dexp,args=list(rate=.2),color="#00FF00")
```

Superimposed Histograms

To create this graph, we first move ourSimulations into a dataframe and then melt that (using the reshape package) to a long format. From there, we are able to make each experiment a very nearly transparent layer (alpha = .007). This allows you to see all experiments at once. We overlay these layers vertical lines to represent the mean and the boundaries of the first standard deviations. We also add a normal curve and an exponential curve so that we can compare those with our data.

```
ourDF <- data.frame(ourSimulations)
colnames(ourDF) <- paste0("X",c(1:1000))
meltDF <- melt(ourDF)
superHist <- ggplot(meltDF)+aes(x=value)
for(i in 1:1000){
  layerNumber <- paste0("X",i)
  superHist <- superHist +
    geom_histogram(aes(y=..density..), data=filter(meltDF,variable==layerNumber), alpha=.007,binwidth=1
}
superHist <- superHist +
  geom_vline(xintercept=mean(meltDF$value), color="#8888FF")+
  geom_vline(xintercept=mean(meltDF$value)-sd(meltDF$value), color="#000088")+
  geom_vline(xintercept=mean(meltDF$value)+sd(meltDF$value), color="#000088")+
  stat_function(fun=dnorm,args=list(mean=5, sd=5), color="#FF00FF")+
  stat_function(fun=dexp,args=list(rate=.2),color="#00FF00")+
  xlab("Values")+ylab("Density")
superHist
```

Histogram of the Means

To create this histogram, we take the ourSimMeans vector and plot the histogram. Again, we add lines showing density, the mean, the range of the first standard deviation, and plots of the exponential and normal curves for comparison.

```
ggplot()+
  aes(ourSimMeans)+
  geom_histogram(aes(y=..density..), alpha=.8, binwidth=.1)+
  geom_vline(xintercept=mean(ourSimMeans), color="#FF8888")+
  geom_vline(xintercept=mean(ourSimMeans)+sd(ourSimMeans), color="#FF0000")+
  geom_vline(xintercept=mean(ourSimMeans)-sd(ourSimMeans), color="#FF0000")+
  geom_density(col="black")+
  stat_function(fun=dnorm,args=list(mean=5, sd=.625), color="#FF00FF")+
  geom_vline(xintercept=5, color="#880088",linetype="dotted")+
  geom_vline(xintercept=5+sqrt(25/40), color="#CC00CC",linetype="dotted")+
  geom_vline(xintercept=5-sqrt(25/40), color="#CC00CC",linetype="dotted")+
  stat_function(fun=dexp,args=list(rate=.2),color="#00FF00")+
  xlab("Sample Means")+ylab("Density")
```

Summary Table

This code was used to create the final summary table

```
theoreticalStatistics <- c(5,25,5)
populationStatistics <- round(c(mean(meltDF$value), var(meltDF$value),sd(meltDF$value)),2)
theoreticalMeansStatistics <- c(5,25/40,sqrt(25/40))
meansStatistics <- round(c(mean(ourSimMeans),var(ourSimMeans),sd(ourSimMeans)),2)
compareStatistics <- rbind(theoreticalStatistics,populationStatistics,theoreticalMeansStatistics, meansS
colnames(compareStatistics) <- c("Mean","Variance","Standard Deviation")
compareStatistics
```