Demonstration of Central Limit Theorem using the Exponential Distribution

Sarah Massengill

1/29/2021

Overview

This Course Project has been designed to demonstrate a student's understanding of how the Central Limit Theorem help to describe the behavior of the distribution of samples of size n. In this report I will demonstrate that the collection of sample means will approach a normal distribution with mean equal to the population mean and variance equal to the population variance divided by the sample size.

To demonstrate the above, I have used the exponential distribution as required for the assignment. Since it was not covered in this course I will fist give a quick over view of the exponential distribution and then move into simulating 1,000 samples of size 40. With the help of graphs I will show that the distribution of means centers around the theoretical mean of an exponential distribution.

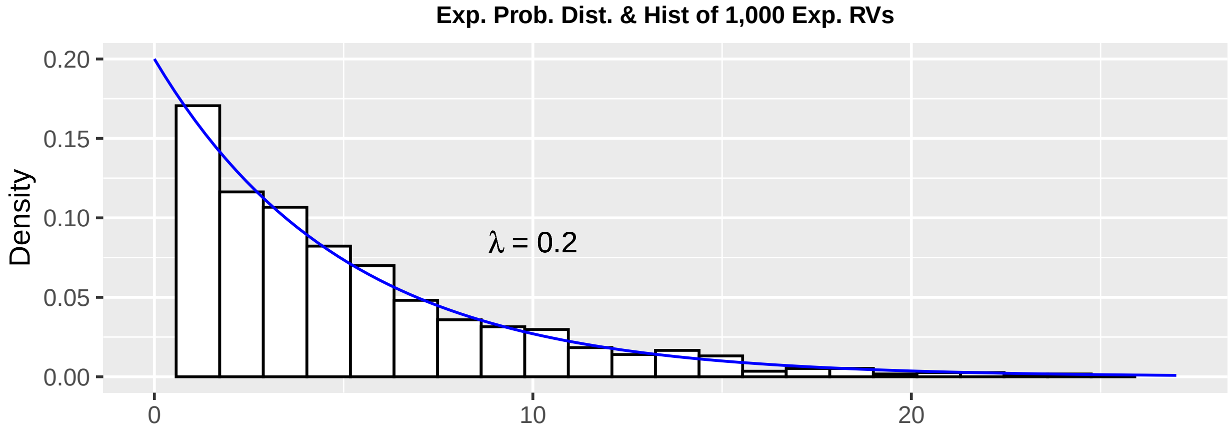
The Exponential Distibution

The exponential distribution is a continuous distribution used to determine the probability that x units of time passes between events. Like the wait time between customers in a bank or a grocery story. The parameter for the exponential function is λ and it represents the number of events that occur in a unit of time. The probability density function for the exponential function is as follows:

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ \lambda e^{-\lambda x} & \text{if } x >= 0 \end{cases}$$

The theoretical mean and variance are as follows:$\mu = \frac{1}{\lambda}$ and $\sigma^2 = \frac{1}{\lambda^2}$

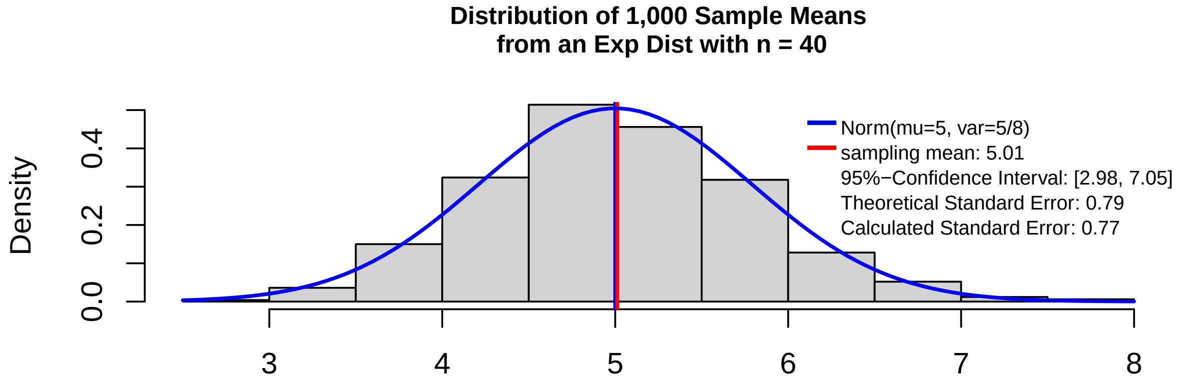
For this project $\lambda = 0.2$ which means that theoretical mean is $\mu = 5$, and variance is $\sigma^2 = 25$. Below is a histogram of a sample of 1,000 random variables selected from an exponential distribution, and sketched on top is the curve for the exponential probability distribution.



Sampling Distribution Mean

In R rexp(40,0.2) can be used to build a sample of $n = 40$ exponential random variables with $\lambda = 0.2$. To simulate a sampling distribution, this is done 1000 times and the means are recorded and a density histogram is used to display the distribution below. From the central limit theorem, we know that this distribution will be normal *not* exponential like the distribution the means were calculated from. This normal distribution of sample means will theoretically have a mean of 5 and a standard error of $25/\sqrt{40} \approx 0.79$. In this case we are using a known population statistic to describe a sampling distribution. Using a 95% confidence interval, the sampling distribution's mean should fall within about 1.65 standard errors of 5. This means that the sampling distributions mean should be within the interval $[2.98, 7.05]$.

The graph below shows that this is in fact what has happened in our simulation of 1,000 samples of size 40. The code for this simulation and the graphs can be found in the appendix.



It is important to note that the population statistics are not normally known and a sample's mean and standard error is used to estimate the population's mean instead of the other way around. If a sample mean was calculated for n observations, a $(1 - \alpha) * 100\%$ confidence interval can be defined to allow the researcher to be $(1 - \alpha) * 100\%$ confident that the population mean is within that confidence interval. Above, $\alpha = .05$ so a 95% confidence interval given for the Sample Means Distribution.

Another thing that is important to note is that the distribution of 1000 sample means is normal with mean very close to 5 (the mean of the exponentially distributed population it was chosen from) and standard deviation equal to the population standard deviation divided by the sample size, $sd = \frac{\sigma}{\sqrt{n}}$. This is not the same as a distribution of 1000 random variables chosen from an exponentially distributed population. These 1000 random variables would be distributed like figure 1.

Appendix

Code:

To create the the exponential distribution curve on top of the 1000 random variables generated from an exponential distribution:

```r
library(ggplot2)

set.seed(1234)
## get value of the 99.5th percentile
xmax <- qexp(0.999, rate=0.2)

## choose equally spaced x-values
xvals <- seq(0, xmax, length=1000)

# create a data frame with 1000 randomly
# generated exponential values to plot
# with the
expdist <- data.frame(values=rexp(1000,rate=0.2))
g<-ggplot(expdist, aes(x=values)) + xlim(0,27)+
   geom_histogram(aes(y =..density..),
                  binwidth=xmax/30,
                  colour="black",
                  fill="white",
                  ) +
  stat_function(fun = dexp, args = list(rate=0.2),colour="blue") +
  annotate(geom="text", x=10, y=0.085,
           label=eval(lambda ~ "= 0.2"),
           color="black") +
  ggtitle("Exp. Prob. Dist. & Hist of 1,000 Exp. RVs") +
  theme(plot.title = element_text(color='black',
                                   size=9,
                                   face="bold",
                                   hjust = 0.5),
        axis.title.x=element_blank())

g
```

To create the sample distribution of means:

```r
set.seed(123)
smu<-NULL
svar<-NULL
n=40
for (i in 1:1000){
  sample = rexp(n,rate = 0.2)
  cmu<-mean(sample)
  cvar<-var(sample)

  smu<-c(smu,cmu)
  svar<-c(svar,cvar)
}
# plot the histogram with output density instead of count
hist(smu,prob=T, ylim=c(0,0.5),breaks=15,
     main = "Distribution of 1,000 Sample Means\n from an Exp Dist with n = 40",
     cex.main = .8,
     xlab = NULL)

# calculate the mean of the sample means
samplingmu = mean(smu)
abline(v=5, col="blue",lwd=2)
abline(v = samplingmu, col="red",lwd=2)

curve(dnorm(x,mean=5, sd=5/sqrt(40)),col="blue", add=T,lwd=2)

conf.int<- samplingmu + c(-1,1)*qnorm(0.995)*5/sqrt(40)

legend("topright",
  legend = c("Norm(mu=5, var=5/8)",
             paste("sampling mean:",round(samplingmu,2)),
             paste("95%-Confidence Interval: ",
                   "[",
                   round(conf.int[1],2),
                   ", ",
                   round(conf.int[2],2),"]"
                            ,sep=""),
             paste("Theoretical Standard Error: ",
                   round(5/sqrt(40),2),sep=""),
             paste("Calculated Standard Error: ",
                   round(sd(smu),2),sep="")),
  col = c("blue","red", "0","0","0"),
  pch = "-",
  bty = "n",
  pt.cex = 2,
  cex = .7,
  text.col = "black",
  horiz = F ,
  inset = c(-0.01, 0)
  )
```