The Exponential Distribution and the Central Limit Theorem; a Brief Exploration

Carlos Schuler

Overview:

In this brief report I explore the behavior of averages of samples from the exponential distribution (using R) and compare their behavior with what we expect from the Central Limit Theorem. The exponential distribution is characterized by the probability density function $f(x) = \lambda e^{-\lambda x}$ for $x > 0$ and $\lambda > 0$. It can be simulated in R with the function *rexp(n, lambda)* where *lambda* is the rate parameter. The mean of the exponential distribution is $\frac{1}{\lambda}$ and the standard deviation is also $\frac{1}{\lambda}$. In this exploration I will focus on $\lambda = 0.2$ for all of the simulations and I will explore the distribution of averages 1000 simulations of 40 samples from this distribution.

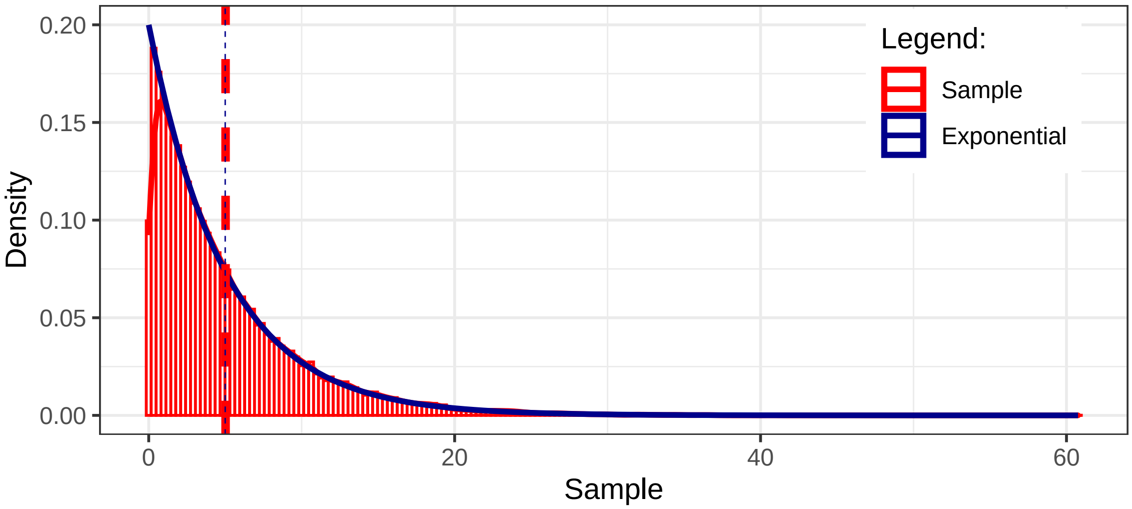
Simulation:

40,000 samples are simulated from the exponential distribution. The figure below shows the histogram for these samples, compared to the theoretical probability density function.

```r
#Initialization and load packages
set.seed(1214); library(ggplot2); library(dplyr)
# Parameters
lambda <- 0.2; sampleSize <- 40; numSimulations <- 1000
sample <- rexp(n= sampleSize*numSimulations, rate= lambda) #<-- The Sample!
theoreticalMean <- 1/lambda; sampleMean = mean(sample)
bw <- 2 * IQR(sample) / length(sample)^(1/3) # Freedman-Diaconis rule for histogram binwidth
```

[ggplot code shown in the appendix]

Fig. 1: Exponential Distribution Sample



Sample Mean versus Theoretical Mean

```r
deltaMeanPercent <- abs(sampleMean-theoreticalMean)/theoreticalMean*100
```

The vertical dashed lines on the plot show the very close agreement between the mean of the sample 5.025 (in red) and that of the population (5) (in blue) - they differ by 0.5%.

To explore the properties of the distribution of the means of 40 samples from the exponential distribution, the 40,000 simulated sample is subdivided into 1000 sets of 40 samples each. The mean of each of the 40-sample sets is calculated, and the "six number summary" for the 1000-long set of 40-sample means is shown below.

```
# sampleMatrix contains numSimulations rows each with sampleSize samples
sampleMatrix <- matrix(sample, nrow=numSimulations,ncol=sampleSize) # <- Split Sample
sampleMeans <- apply(sampleMatrix,1,mean) # <- Means for each 'sampleSize' sample
sampleSummary<-summary(sampleMeans); meanSampleMeans <- sampleSummary["Mean"]
maxSampleMeans <- sampleSummary["Max."]; minSampleMeans <- sampleSummary["Min."]
varSampleMeans <- var(sampleMeans); theoreticalVar <- 1/lambda^2/sampleSize
sampleSummary
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.576   4.464   4.979   5.025   5.523   7.789
```

Sample Variance versus Theoretical Variance

As expected, the "mean of the 40-sample means" coincides with the mean of the original large sample, already discussed above. The calculated "variance of the 40-sample means" is 0.6396, which agrees reasonably well with the expected value of $1/N\lambda^2 = 0.625$, the difference between the sample estimate and the population variance is 2.34%.

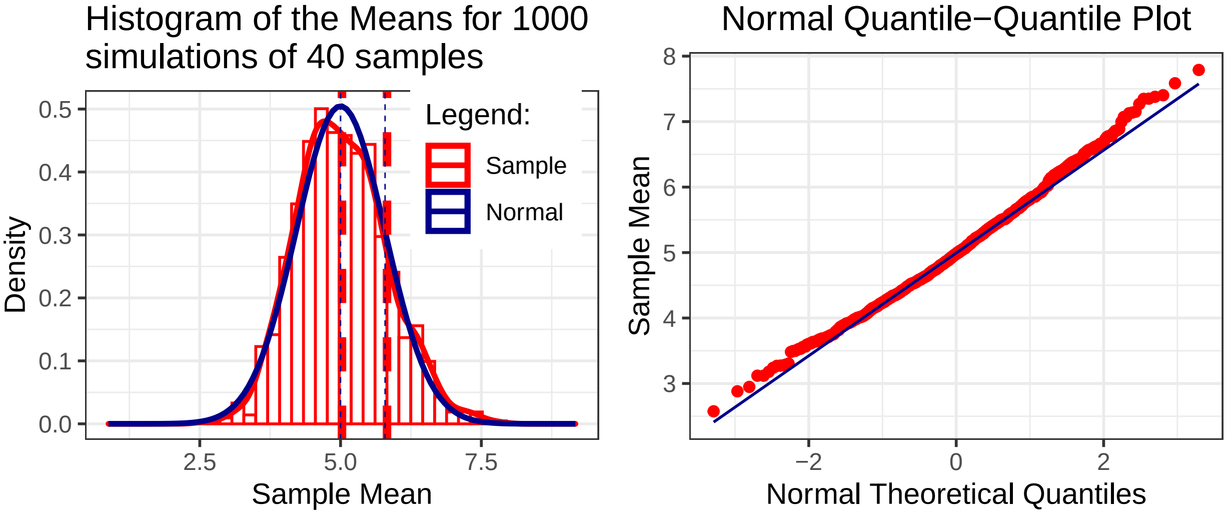
Distribution

The left panel in the figure below shows the histogram for the calculated 40-sample means, and compares it to the expected theoretical normal distribution with a mean of $1/\lambda = 5$ and a variance of $1/N\lambda^2 = 0.625$. The right panel shows the Quantile-Quantile (Q-Q) normal probability plot for the calculated 40-sample means.

```
# Calculate xlims to make the plot symmetric
xmax <- max(maxSampleMeans-meanSampleMeans,meanSampleMeans-minSampleMeans)
xAxisMin <- meanSampleMeans-1.5*xmax; xAxisMax <- meanSampleMeans+1.5*xmax
bw <- 2 * IQR(sampleMeans) / length(sampleMeans)^(1/3) # Freedman-Diaconis rule
sampleMeansDF<- data.frame(means=sampleMeans)
```

[ggplot code shown in the appendix]

Fig. 2: Comparison wih the Normal Distribution



In the histogram plot, the red and blue vertical lines in the center shows the overlap between the "mean of the 40-sample means" and the expected population value of $1/\lambda$. The vertical line to the right was drawn 1 standard deviation to the right (calculated independently for the sample and the population) and the red/blue overlap shows the close agreement between the calculated sample variance and the population value $1/N\lambda^2$.

Both the histogram representation (overlapping of the sample and population density functions), as well as the Q-Q normal plot (overlap of the data with the diagonal straight line) show that the distribution of the 40-sample means is well approximated by a normal distribution.

The distribution of a large collection of random samples from an exponential distribution is quite different from the distribution of a large collection of averages of 40 samples from the same distribution.

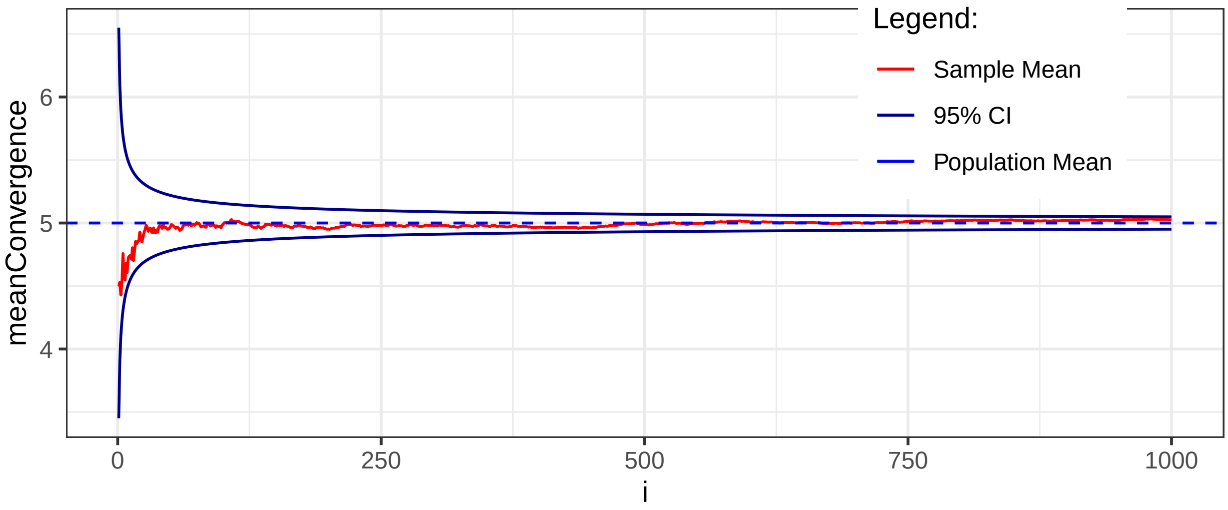
Convergence of the "Mean of 40-Sample Means" and Confidence Intervals

The figure below shows the convergence of the "mean of 40-sample means" as the number of simulations goes from 1 to 1000. In the figure, the mean is compared to the 95% confidence interval(dark blue dashed lines), assuming the distribution of the 40-sample means is normal. The plot shows good coverage of the confidence interval for the calculated sample mean value as a function of the number of simulations.

```
meanConvergence <- cumsum(sampleMeans)/1:numSimulations
upperLim <- qnorm(0.975,mean=theoreticalMean,sd=sqrt(theoreticalVar/1:numSimulations))
lowerLim <- qnorm(0.025,mean=theoreticalMean,sd=sqrt(theoreticalVar/1:numSimulations))
convergenceDF <-data.frame(i=1:numSimulations,meanConvergence=meanConvergence,
                           upperLim=upperLim,lowerLim=lowerLim)
```

[ggplot code shown in the appendix]

Fig.3 Convergence of the Mean



Conclusion

These simulations show that the distribution of averages of 40-samples from the exponential distribution reasonably approximate the normal distribution with the mean and variance predicted by the Central Limit Theorem.

APPENDIX: ggplot code for the figures

Fig. 1

```
suppressWarnings(print(data.frame(sample = sample) %>% ggplot(., aes(sample)) +
        geom_histogram(aes(y=..density..,color="Sample"), position="identity",
                       fill="white", binwidth=bw) + geom_density(aes(color="Sample"),size=1) +
        stat_function(fun = dexp, aes(color="Exponential"), size=1,args = list(rate=lambda)) +
        scale_color_manual("Legend:",breaks=c("Sample","Exponential"),values=c("red","darkblue")) +
        ggtitle ("Fig. 1: Exponential Distribution Sample") + xlab("Sample") + ylab("Density") +
        geom_vline(xintercept=sampleMean,size=1.5,color="red",linetype="dashed") +
        geom_vline(xintercept=theoreticalMean,size=0.25,color="darkblue",linetype="dashed") +
        theme_bw() + theme(legend.position = c(0.85,0.8))))
```

Fig. 2

```
suppressWarnings({
plot1 <- ggplot(sampleMeansDF, aes(means)) + geom_histogram(aes(y=..density..,color="Sample"),
        position="identity", fill="white", binwidth=bw) +geom_density(aes(color="Sample"),
        size=1) + stat_function(fun = dnorm, aes(color = "Normal"), size=1,
        args = list(mean = theoreticalMean, sd = sqrt(theoreticalVar))) +
  scale_color_manual("Legend:", breaks=c("Sample","Normal"),values=c("red","darkblue")) +
  ggtitle (paste0("Histogram of the Means for ", numSimulations,"\nsimulations of ",sampleSize,
        " samples")) + xlab("Sample Mean") + ylab("Density") +
  geom_vline(xintercept=meanSampleMeans,size=1.5,color="red",linetype="dashed") +
  geom_vline(xintercept=theoreticalMean,size=0.25,color="darkblue",linetype="dashed") +
  geom_vline(xintercept=meanSampleMeans+sqrt(varSampleMeans),size=1.5,color="red",linetype="dashed") +
  geom_vline(xintercept=theoreticalMean+sqrt(theoreticalVar),size=0.25,color="darkblue",
  linetype="dashed") + xlim(xAxisMin,xAxisMax)+theme_bw()+theme(legend.position = c(0.8,0.75))
plot2 <- ggplot(sampleMeansDF, aes(sample=means))+ stat_qq(distribution=qnorm, color="red") +
  stat_qq_line(distribution=qnorm,color="darkblue") + ggtitle("Normal Quantile-Quantile Plot") +
  ylab("Sample Mean") + xlab("Normal Theoretical Quantiles") + theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
gridExtra::grid.arrange(plot1,plot2,ncol=2,
                        top = "Fig. 2: Comparison wih the Normal Distribution")})
```

Fig. 3

```
ggplot(convergenceDF) + geom_line(aes(x=i,y=meanConvergence,color="Sample Mean")) +
  geom_line(aes(x=i,y=upperLim,color="95% CI")) + geom_line(aes(x=i,y=lowerLim,color="95% CI"))+
  geom_hline(aes(yintercept=theoreticalMean, color="Population Mean"),linetype="dashed") +
  scale_color_manual("Legend:", breaks=c("Sample Mean","95% CI","Population Mean"),
  values=c("red","darkblue","blue")) + theme_bw() + theme(legend.position = c(0.8,0.8)) +
   ggtitle("Fig.3 Convergence of the Mean") +
  theme(plot.title = element_text(hjust = 0.5))
```