Statistical Inference Part 1 - Central Limit Theorem

Betsy Nash

January 13, 2018

Synopsis

The question posed is to validate the Central Limit Theorem using simulated data from the exponential distribution.

CENTRAL LIMIT THEOREM (CLT) - This is a material statistical theorem.
The CLT states that under certain conditions, the averages (or the means) of the variables resembles a standard normal distribution as the sample size increases. Those conditions are: 1) the variables are independent, and 2) the variables are identically distributed.

The result of this review is that the CLT has been validated using simulated exponential variables.

Create the DataSet

The exponential distribution is defined with a mean and a variance.
The mean = 1/lambda and the variance = 1/(lambda^2).
The lambda parameter is given in the assignment to be 0.2. The assignment also states the sample size to be 1,000 estimates of the mean, where the mean is calculated over 40 exponential variables. Therefore 40,000 exponential variables are needed to generate our desired 1,000 x 40 dataset matrix.

Under CLT, the theoretical mean for N observations is also 1/lambda. The theoretical variance is [(1/lambda^2)/N]. N in this case is 40.

```
lambda<-0.2
#theoretical mean
meanE<-1/lambda
#theoretical variance CLT
N<-40
VarE<-((1/(lambda^2))/N)
#theoretical standard deviation CLT
stdevE<-sqrt(VarE)
#matrix with 40,000 simulated values: 1,000 rows, 40 columns
mymatrix<-matrix(rexp(40000,lambda),1000,40)
```

Now that the matrix is availble, calculate the mean of each of the 1,000 rows. First the data will be explored for reasonability. Then the data will be validated to see of the CLT holds true. Validation Test: This data should resemble a standard normal distribution under the CLT assumption.

```
myrowmean<-apply(mymatrix,1,mean)
```

Explore the 1,000 Calculated Means

The 5-number summary is used to check for reasonability at a high level.
In addition, a comparison of the expected theoretical mean and variance will be compared to the simulated sample statistics.

```
#five number summary (min, max, etc)
summary(myrowmean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.570   4.428   4.978   4.989   5.492   7.970
```

```
#sample mean
mean(myrowmean)
```

```
## [1] 4.988905
```

```
#sample variance (function uses n-1, convert to n)
var(myrowmean)*(40-1)/40
```

```
## [1] 0.6535236
```

```
#sample std dev (function uses n-1, convert to n)
sd(myrowmean)*(40-1)/40
```

```
## [1] 0.798239
```

No unusual results (ie negative or extreme values) are present in the five number summary. The reasonability test passes.

Comparisons-

Mean: The theoretical mean is 5 and the simulated sample mean is 4.9889054. The means are quite close in magnitude.

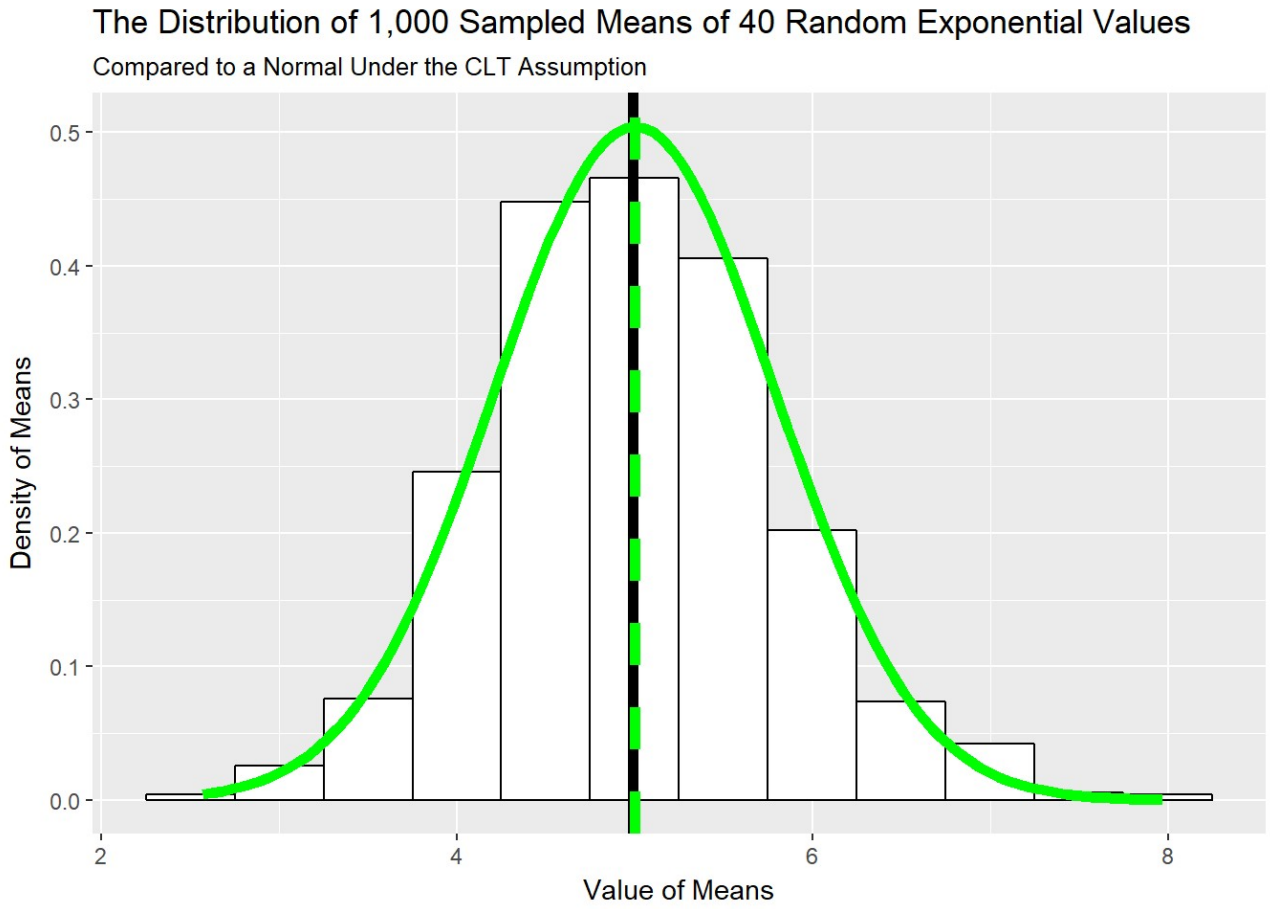
Variance: The theoretical variance is 0.625 and the simulated sample variance is 0.6535236. Deviations can be expected given the nature of sampling. The numbers are reasonably close in magnitude.

StdDev: The theoretical standard deviation is 0.791 and the simulated sample standard deviation is 0.798239.Likewise, deviations can be expected for random samples. The numbers are close in magnitude. The next step is to chart the means of the sampled data to see if the shape resembles a standard normal distribution.

Plot the Data of 1,000 Means

Provided the CLT holds, the curve should be bell-shaped. In addition, a fitted normal distribution will be overlayed for reference.

```
myplot<- ggplot(data.frame(myrowmean = myrowmean), aes(x = myrowmean))
myplot = myplot + geom_histogram(binwidth=0.5, color = "black", fill = "white", aes(y
=..density..))
#labels
myplot = myplot + ggtitle("The Distribution of 1,000 Sampled Means of 40 Random Expone
ntial Values", subtitle = "Compared to a Normal Under the CLT Assumption")+ xlab("Valu
e of Means") + ylab("Density of Means")
#sample mean
myplot = myplot + geom_vline(xintercept = mean(myrowmean), size = 2)
#theoretical mean
myplot = myplot + geom_vline(linetype="dashed",xintercept = as.numeric((meanE)), size
= 2, col="green")
#normal curve using CLT and standard deviation = sigma/sqrt(n), where n = number of es
timates used to calculate the mean
myplot = myplot + stat_function(fun=dnorm, geom = "line",
              args = list(mean = meanE, sd = stdevE),
              size = 2, col = "green")
myplot
```



The curve is bell-shaped with closely-aligned means.

Conclusion

This study validates the CLT using simulated exponential variables. The distribution of the means resembles a standard normal distribution.