Coursera Statistical Inference Class Project - Part I - Exponential Distribution, Simulations and Analysis

CRR

Saturday, June 20, 2015

OVERVIEW: In this part of the project, the exponential distribution is investigated using R and compared with the Central Limit Theorem - a distribution of averages of 40 exponentials is used. The specific goals for the project are:

1. Show the sample mean and compare it to the theoretical mean of the distribution (of averages).
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution (of averages).
3. Show that the distribution (of averages) is approximately normal versus the distribution of a large collection of random exponentials.

SIMULATIONS: The simulations done were written in R code and are shown in the appendix below. Comments within the code itself explain what is being done in each section. Basically, a distribution of 1000 sets of averages of 40 exponentials (lambda or rate=0.2) was generated and analysed; this distribution was compared with a distribution of 1000 random exponentials (lambda or rate=0.2) that was generated as well.

RESULTS:

Means. The sample mean (5.04) and the theoretical mean (1/lambda = 5.0) for the distribution of averages are close - they are illustrated by the blue and red vertical lines (respectively) in the Histogram of Averages figure given in the Appendix below. The sample mean represents the center of the distribution of averages as generated and the theoretical mean the center of the distribution of averages as predicted by the Central Lmit Theorem.

Variances/Variabilities. The sample variance (0.615) and the theoretical variance (1/(lambda*lambda)/n = 0.625; n=40) of the distribution of averages are not very large and are close in size. They represent variability and are illustrated by the widths of the blue and red normalized curves (respectively) shown in the Histogram of Averages figure given in the Appendix. The Central Limit Theorem predicts that the difference in these variances would decrease as the number of averages considered increased.

Distribution. The Histogram of Averages figure (see Appendix below) shows the distribution of the averages of 1000 sets of 40 random exponentials generated. The sample and theoretical means are indicated by the blue and red vertical lines, respectively. The sample normalized curve overlay (in blue) is derived using the sample mean and standard deviation (square root of the variance); the theoretical normalized curve overlay (in red) is derived from using the corresponding theoretical values. The sample mean indicates the center of the distribution of the averages of the exponentials, as well as the center of the sample normalized curve; the theoretical mean indicates the center of the theoretical normalized curve.

The sample and theoretical means are very close to each other as are the normalized curves - the latter (considering their widths) indicate that the variances (variability) are not large and are very close as well. The histogram itself follows closely to the sample normalized curve which indicates that the averages of the exponentials are normally distributed (as predicted by the Central Limit Theorem) - this is in contrast to the distribution of random exponentials as shown in the Histogram of Exponentials that follows in the Appendix.

The applicability of the Central Limit Theorem is demonstrated - the distribution of averages is (nearly) normal with mean=mu and variance=sigma squared/n. The slight deviations noted would be expected to disappear if the number of averages considered was increased above 1000.

See Appendix below.

APPENDIX : R code (with comments) and the values calculated and figures produced by it.

```
## Exponential Distribution Investigation

## Setting seed for randomization
set.seed(575)

## Generating means of samples of 40 exponentials - 1000 means result
## using lambda or rate =0.2 for the exponentials
mns=NULL
for (i in 1:1000) mns=c(mns,mean(rexp(n=40, rate=0.2)))

##Sample mean is mean(mns) and Theoretical mean is (1/rate)
print ("Sample Mean =");mean(mns)
```

```
## [1] "Sample Mean ="

## [1] 5.03909
```

```
print ("Theoretical Mean =");1.0/0.20
```

```
## [1] "Theoretical Mean ="

## [1] 5
```

```
##Sample variance is var(mns) and Theoretical variance is (1/rate)squared divided by n
print ("Sample Variance =");var(mns)
```

```
## [1] "Sample Variance ="

## [1] 0.6146269
```

```
print ("Theoretical Variance =");(1.0/0.20)*(1.0/0.20)/40
```

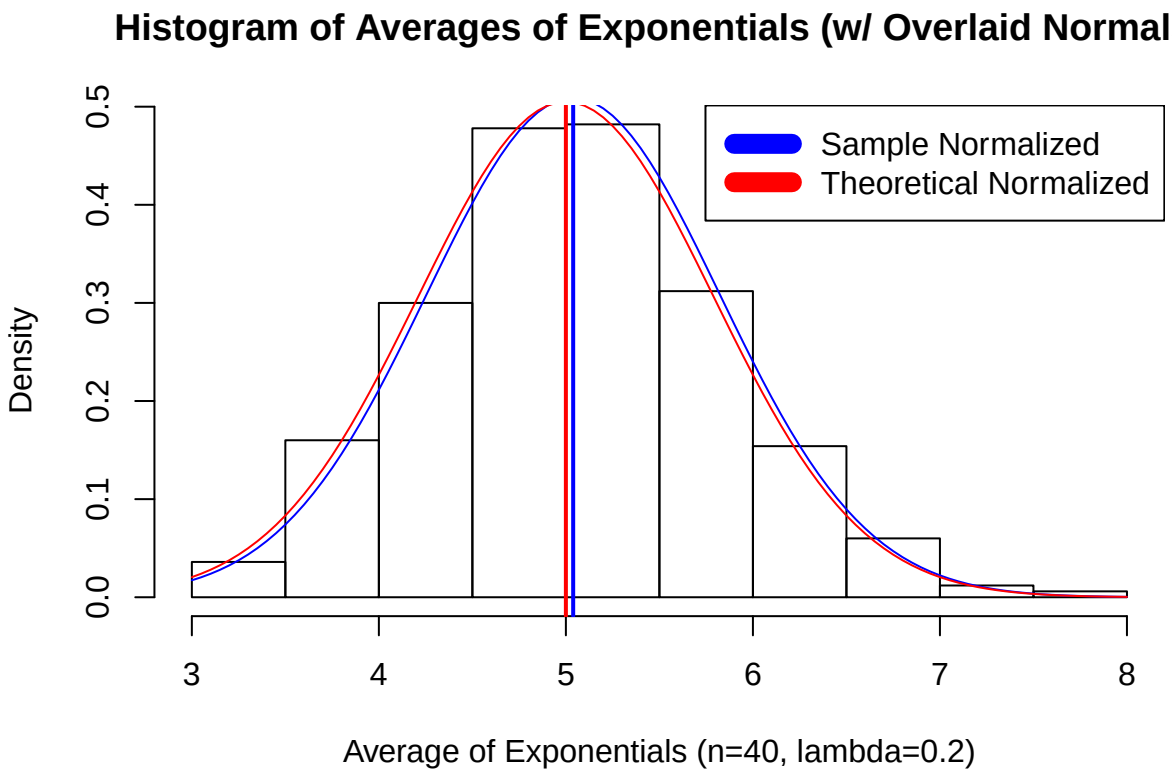
```
## [1] "Theoretical Variance ="

## [1] 0.625
```

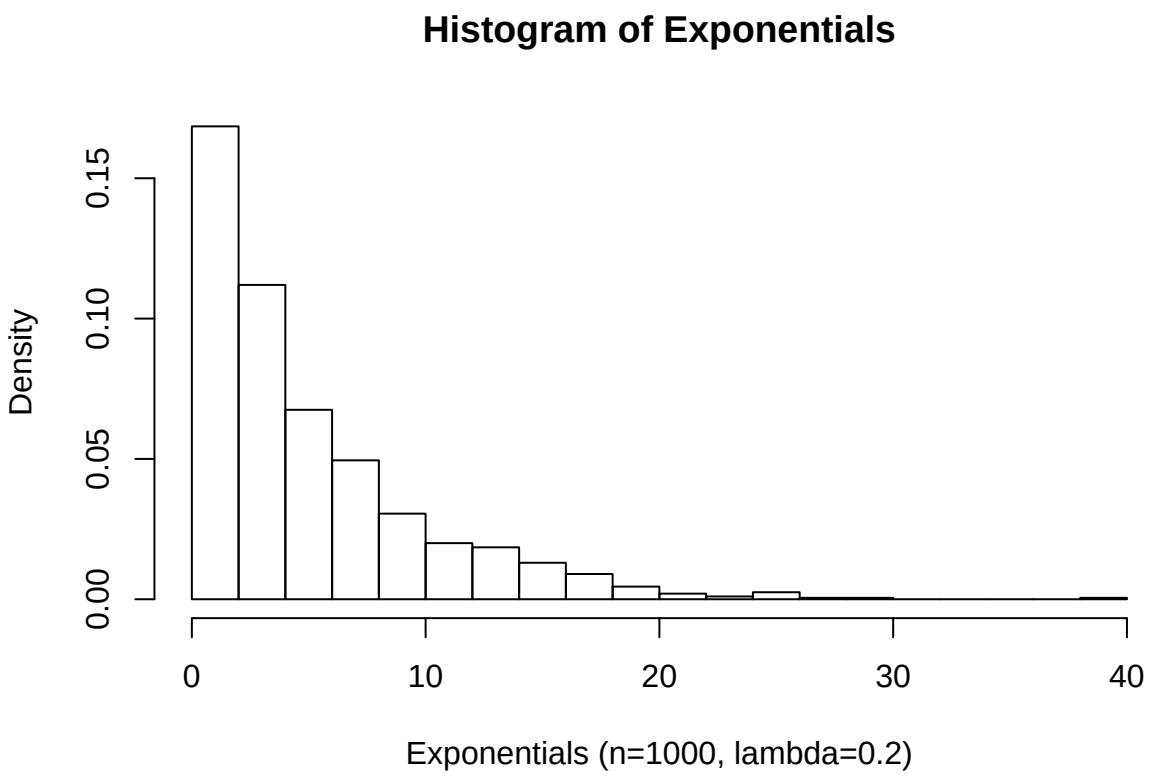
```
##Theoretical mean is (1/rate) = 5.0 and theoretical standard deviation is square root of theoretical va

## Histogram Plot of Averages
hist(mns, prob=TRUE, main="Histogram of Averages of Exponentials (w/ Overlaid Normals)", xlab="Average
## Sample mean on plot
abline(v = mean(mns), col = "blue", lwd = 2)
## Theoretical mean on plot
abline(v = 5.0, col = "red", lwd = 2)
## Normal curve for plot with sample mean and standard deviation
curve(dnorm(x, mean=mean(mns), sd=sd(mns)), col="blue", add=TRUE)
## Normal curve for plot with theoretical mean and standard deviation
curve(dnorm(x, mean=5.0, sd=0.79), col="red", add=TRUE)

legend("topright", c("Sample Normalized", "Theoretical Normalized"), col=c("blue", "red"), lwd=10)
```




```
## Histogram Plot of Exponentials for Comparison
set.seed(575)
hist(rexp(n=1000, rate=0.2), prob=TRUE, breaks=20, main="Histogram of Exponentials", xlab="Exponentials
```


Histogram of Exponentials
Density
0.00
0.05
0.10
0.15
0
10
20
30
40
Exponentials (n=1000, lambda=0.2)