Statistical Inference Course Project

Darrell Nabors

March 19, 2015

In this project you will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. ***Set lambda = 0.2 for all of the simulations.*** You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. You should 1. Show the sample mean and compare it to the theoretical mean of the distribution. 2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution. 3. Show that the distribution is approximately normal.

In point 3, focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

First off, we will define the variables *lambda*, *n* and *simulations* as well as call the *ggplot2* library. We will calculate 40 exponentials to be used in the analytical mean calculation.

```
library(ggplot2)
lambda <- 0.2
n <- 40
simulations <- 1000
forty_simulated_exponentials <- replicate(simulations, rexp(n,lambda))
forty_means_exponentials <- apply(forty_simulated_exponentials,2,mean)
```

1. Show the sample mean and compare it to the theoretical mean of the distribution.

```
analytical_mean <- mean(forty_means_exponentials)
theory_mean <- 1/lambda
analytical_mean
```

```
## [1] 4.995336
```

```
theory_mean
```

```
## [1] 5
```

The Analytical mean = 4.9953362, while the Theoretical mean = 5. As expected, the theoretical mean is very sinilar to the analytical mean.

2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.

```
theory_sd <- ((1/lambda)*(1/sqrt(n)))
actual_sd  <-sd(forty_means_exponentials)
theory_var <- theory_sd^2
actual_var <- var(forty_means_exponentials)
theory_sd
```

```
## [1] 0.7905694
```

```
actual_sd
```

```
## [1] 0.7349906
```

```
theory_var
```

```
## [1] 0.625
```

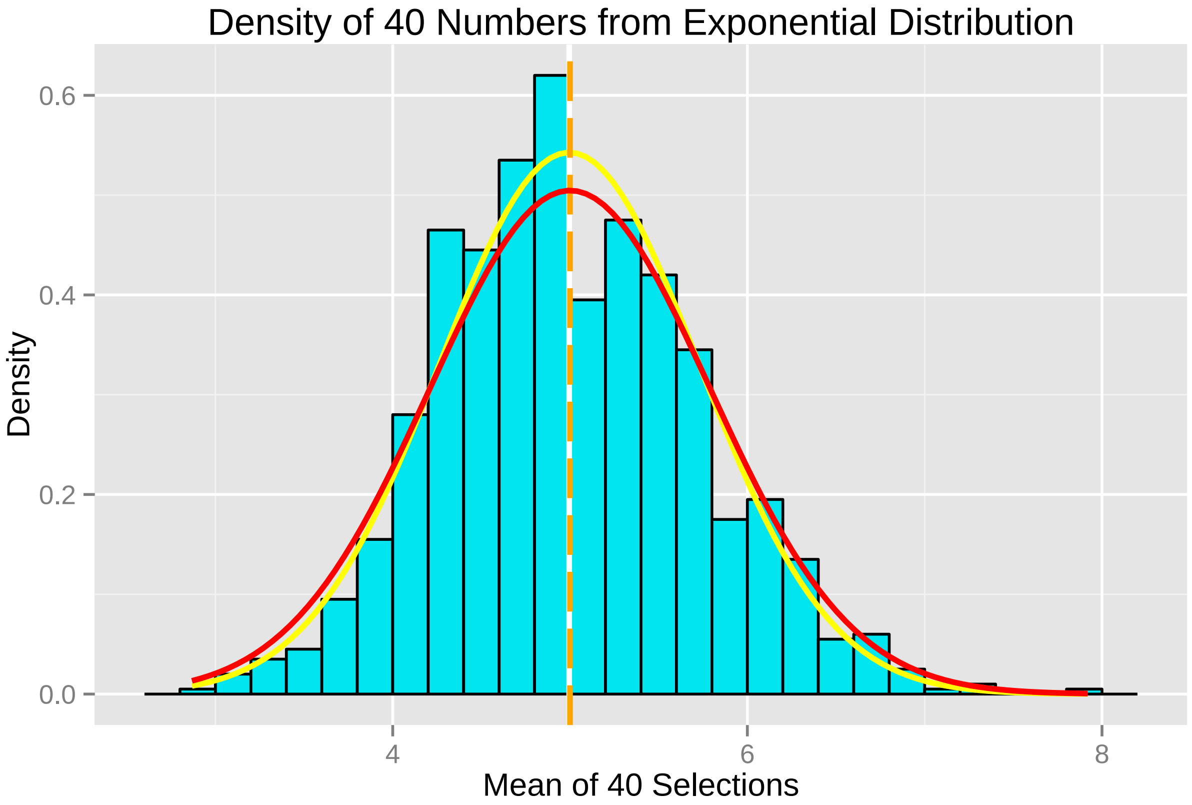
```
actual_var
```

```
## [1] 0.5402111
```

As can be observed, the actual standard deviation (0.7349906) and the theoretical standard deviation (0.7905694) are very similar, while the difference between the actual variance (0.5402111) and the theoretical variance (0.5402111) is slightly greater than the deviation figures, but still less than 1.

3. Show that the distribution is approximately normal.

```
df_forty_means <- data.frame(forty_means_exponentials)
p1 <- ggplot(df_forty_means,aes(x=forty_means_exponentials))
p1 <- p1 +
geom_histogram(binwidth=lambda,fill="turquoise2",color="black",aes(y =..density..))
p1 <- p1 +
labs(title="Density of 40 Numbers from Exponential Distribution",x="Mean of 40 Selections",y="Density")
p1 <- p1 +
geom_vline(xintercept=analytical_mean,size=1.0, color="grey100")
p1 <- p1 +
stat_function(fun=dnorm,args=list(mean=analytical_mean,sd=actual_sd),color="yellow",size=1.0)
p1 <- p1 +
geom_vline(xintercept=theory_mean,size=1.0,color="orange1",linetype="longdash")
p1 <- p1 +
stat_function(fun=dnorm,args=list(mean=theory_mean,sd=theory_sd),color="red",size = 1.0)
p1
```



The distribution of averages of the actual data (pictured in yellow) is very close to a normal distribution. This phenomena can be explained by the Central Limit Theorem (CLT). The CLT states that the distribution of the sum (or average) of a large number of independent, identically distributed variables will be approximately normal, regardless of the underlying distribution. The actual mean is the white dotted line, while the theorical mean is the orange dotted line.