Statistical Inference - The Central Limit Theorem

Ron Ferens

September 26, 2015

Overview

The *Statistical Inference* curse project, investigates the exponential distribution in R and compare it with the Central Limit Theorem (CLT). The CLT states that the distribution of averages of iid variables (properly normalized) becomes that of a standard normal as the sample size increases. The result is that: $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$. as a distribution like that of a standard normal for large n.

Based on CLT is we can approximate as followed: $\bar{X}_n \sim N(\mu, \sigma^2/n)$

The thoretical mean of an exponential distribution is $\mu = \frac{1}{\lambda}$ and the standard deviation is $\sigma = \frac{1}{\lambda}$.

When appling the CLT to the exponential distribution and we can estimate the mean and variance based on the sampled data. In this report, based on the CLT, we will use $n = 40$ observations with $\lambda = 0.2$ to show that the sampling distribution of the mean of an exponential distribution is indeed approximately $N(\frac{1}{0.2}, \frac{\frac{1}{0.2}}{\sqrt{40}})$ distributed.

Sample Mean versus Theoretical Mean

To investigate the exponential distribution in R and compare it with the Central Limit Theorem (CLT) we will repeatedly (1000 iterations) draw $n = 40$ observations of the follwoing exponential distributionwith $\lambda = 0.2$. For each of draw, we will calculate the mean of the n observations. Theoretically, we should get a good estimation of the theoretical mean of the exponential distribution is $\frac{1}{\lambda}$.

```
nosim <- 1000
n <- 40
lambda <- 0.2

## Simulate nosim averages of 40 exponential distribution
mean(apply(matrix( rexp((nosim * n), lambda), nosim), 1, mean))
```

```
## [1] 5.008433
```

```
## Theoretical mean of the exponential distribution is 1/lambda
1/lambda
```

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

As shown above, the estimated mean is pretty close to the theoretical mean of the exponential distribution which is $\mu = \frac{1}{0.2} = 5$

Sample Variance versus Theoretical Variance

In order to compare the sample variance to the theoretical variance we will repeat the process we did for the sample mean (previous section). This time we will calculate the variance of n observations in each draw (total of 1000 draws).

```
## Simulate nosim averages of 40 exponential distribution
mean(apply(matrix( rexp((nosim * n), lambda), nosim), 1, var))
```

```
## [1] 24.99786
```

```
## Theoretical variance of the exponential distribution is lambda^-2
lambda^-2
```

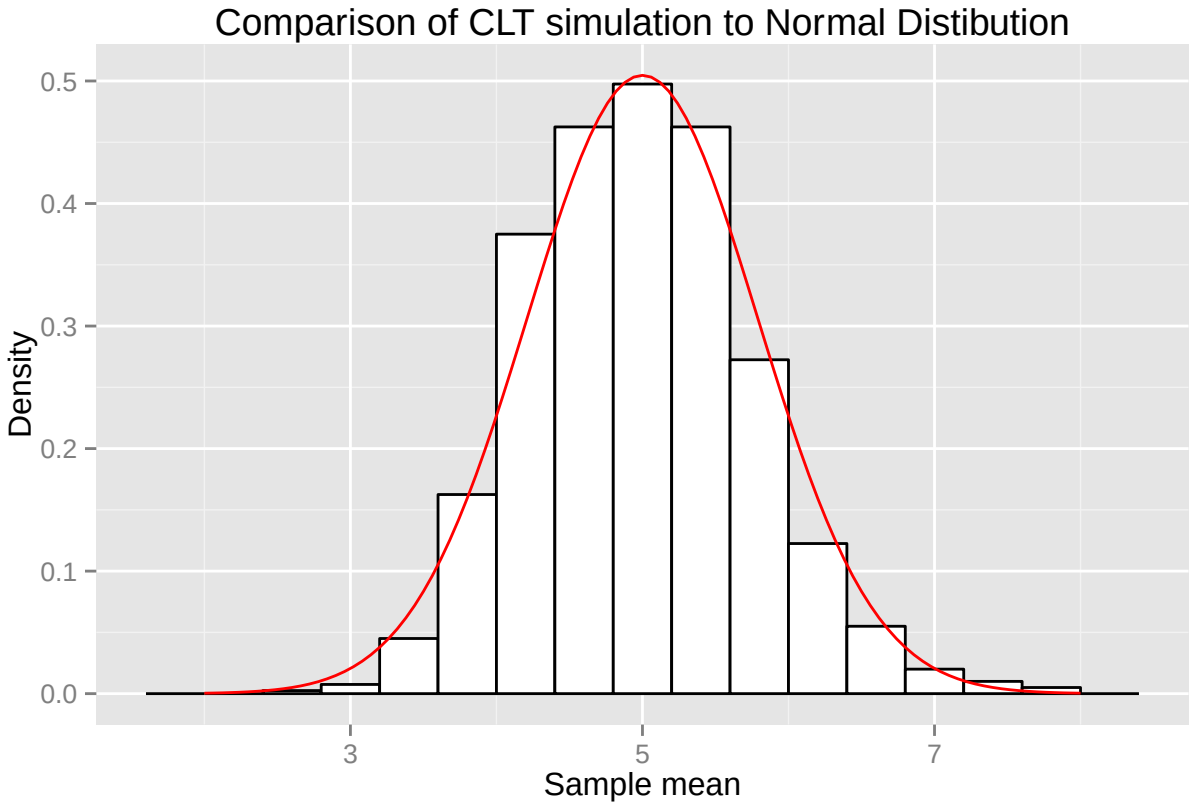
```
## [1] 25
```

As expected, the averaged variance of the simulated data is very close to the theoretical variance of the exponential distribution is $Var = \frac{1}{\lambda^2} = 25$

Distribution

To verify that the distribution of the CLT simulation is approximately normal we will compare it to the following theoretical normal distribution: $N(\frac{1}{0.2}, \frac{\frac{1}{0.2}}{\sqrt{40}})$

```
x <- apply(matrix(rexp((nosim * n), lambda), nosim), 1, mean)
data <- as.data.frame(x)
ggplot(data, aes(x = x)) +
  geom_histogram(binwidth = 0.4, color = 'black', fill = 'white', aes(y = ..density..)) +
  stat_function(aes(x = c(2, 8)), fun = dnorm, color = 'red',
                args = list(mean = 1/lambda, sd = (1/lambda/sqrt(n)))) +
  xlab('Sample mean') +
  ylab('Density') +
  ggtitle('Comparison of CLT simulation to Normal Distibution')
```



in the plot above, it can be easily seen that sample distribution is very similar to the overlayed red normal distribution. This assures us that CLT simulation is approximately normal.