Statistical Inference Project - Part 1

Oye Akinnikawe

January 1, 2017

Synopsis

This is a project for the Cousera Statistical Inference Class. The project consists of two parts:

1. Simulation exercise to explore inference
2. Basic inferential data analysis using the ToothGrowth data in the R datasets package

Part 1 - Simulation Exercise

Overview

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.*

Instructions

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.

Load Libraries

```
library(ggplot2)
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 3.2.5
```

define problem variables

```
n <- 40 ## sample size, number of exponentials
nosim <- 1000 ## number of simulations
lambda <- 0.2 ## rate parameter
set.seed(1234) ## seed value for reproducibility
mu <- 1/lambda ## mean of exponential distribution
sigma <- 1/lambda ## standard deviation of exponential distribution
se <- sigma/sqrt(n) ## standard error
```

create a matrix with 1000 simulations each with 40 samples drawn from the expenetial distribution and find the means.

```
simMeans <- apply(matrix(rexp(nosim * n, lambda), nosim), 1, mean)
```

Mean Comparison

Estimate the sample mean and compare it with the theoretical mean of the distribution.

Sample Mean

The average sample mean of 1000 simulations of 40 randomly sampled exponential distributions.

```
sampleMean <- mean(simMeans)
sampleMean
```

```
## [1] 4.974239
```

Theoretical Mean

The expected mean of the exponential distribution of rate = 1/lambda.

```
theoryMean <- 1/lambda
theoryMean
```

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

the sample mean of the exponential distribution is **4.974239** and the theoretical mean of the distribution is **5**.

Variance Comparison

Estimate the sample variance and compare it with the theoretical variance of the distribution.

Sample Variance

The sample variance of 1000 simulations of 40 randomly sampled exponential distributions.

```
sampleVar <- var(simMeans)
sampleVar
```

```
## [1] 0.5949702
```

Theoretical Variance

the expected variance of the exponential distribution is sigma^2 / n

```
theoryVar <- sigma^2/n
theoryVar
```

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

the sample variance of the exponential distribution is **0.595** and the theoretical variance of the distribution is **0.625**. The difference is **0.03**.

Standard deviation

Estimate the standard deviation of the exponential distribution

```
sampleSD <- sd(simMeans)
theorySD <- sigma/sqrt(n)
```

```
## Sample Std. deviation is:  0.7713431
```

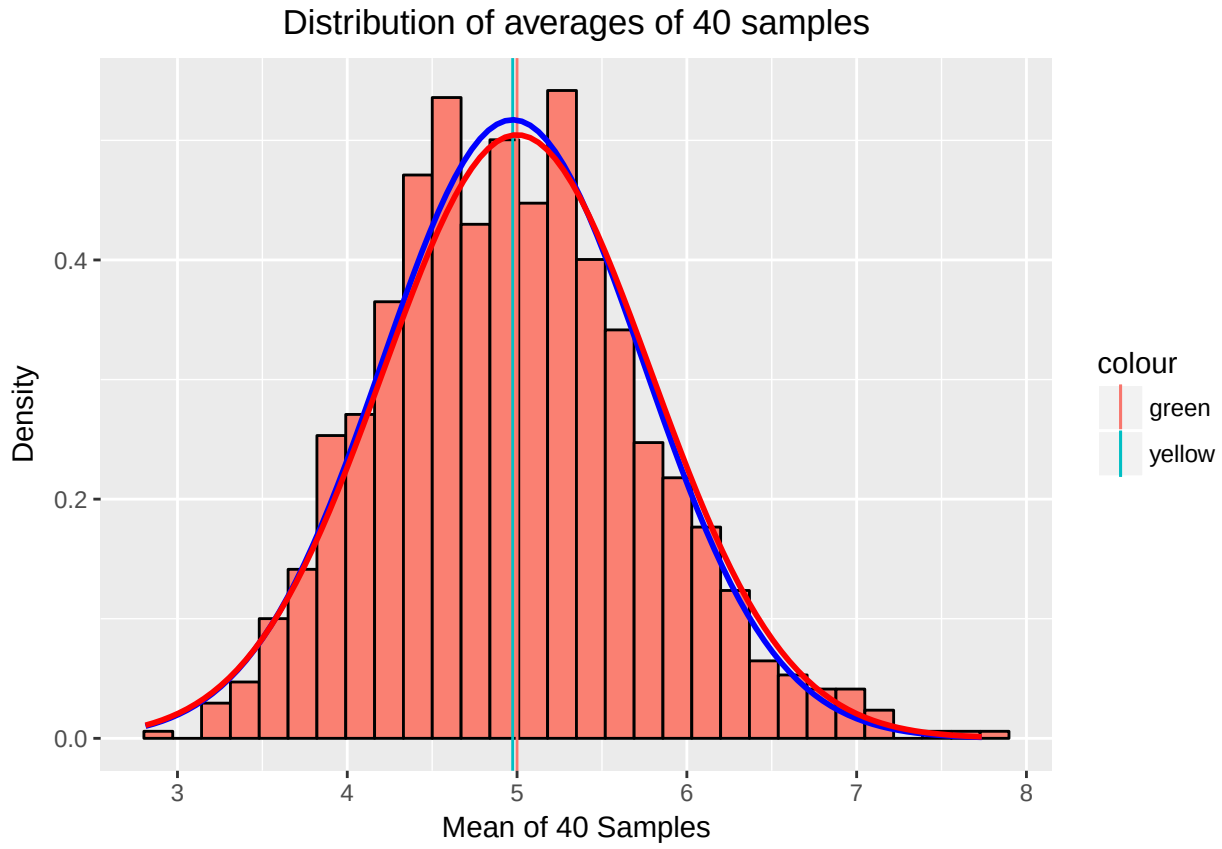
```
## Theoretical Std. deviation is:  0.7905694
```

Show that the distribution is Approximately normal

Display results that visually compares the sample values versus the theorectical values.

```
plotdata <- data.frame(simMeans)
g <- ggplot(plotdata, aes(x = simMeans))
g <- g + geom_histogram(aes(y = ..density..), colour = "black", fill = "salmon")
g <- g + labs(title = "Distribution of averages of 40 samples", x = "Mean of 40 Samples", y = "Density"
g <- g + geom_vline(aes(xintercept = sampleMean, colour = "yellow"))
g <- g + geom_vline(aes(xintercept = theoryMean, colour = "green"))
g <- g + stat_function(fun = dnorm, args = list(mean = sampleMean, sd = sampleSD), colour = "blue", siz
g <- g + stat_function(fun = dnorm, args = list(mean = theoryMean, sd = theorySD), colour = "red", size
g <- g + theme(plot.title = element_text(hjust = 0.5))
g
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The Central limit Theorem (CLT) states averages are approximately normal, with distributions

- centered at the population mean
- with standard deviation equal to the standard error of the mean

We can see that the resulting distribution looks like a bell curve and the standard deviation of 0.77 is approximately equal to the standard error of the mean 0.79.

Confidence Intervals comparison

Estimate the sample confidence intervals (CI) and compare it to the theoretical CI

sample CI

```
sampleCI <- round (mean(simMeans) + c(-1,1) * qnorm(.975) * sd(simMeans) / sqrt(n),3)
sampleCI
```

```
## [1] 4.735 5.213
```

Theoretical CI

```
theoryCI <- theoryMean + c(-1,1) * qnorm(.975) * sqrt(theoryVar)/sqrt(n)
theoryCI
```

```
## [1] 4.755005 5.244995
```

The sample CI and the theoretical CI are a good match

Conclusion

The exponential distribution demonstrates the Central Limit Theorem, a bell shape curve. The 95% confidence interval of the sample and theoretical CI are approximately equal.