Coursera Statistical Inference Course Project Part 1

kimnewzealand

29 May 2017

Overview

In part 1 of this project we are investigating the exponential distribution in R and comparing it with the Central Limit Theorem using a simulation exercise.

Setup

Load packages

For this analysis, the following R packages are needed `ggplot2`.

```
library(ggplot2)
```

Part 1: Simulation Exercise

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

As per the instructions, the exponential distribution can be simulated in R with the function rexp(n, lambda) where lambda is the rate parameter. The simulation can be repeated multiple times using the repetition function.

The theoretical mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda.

For the 1000 simulations, lambda is assumed to be 0.2 and the sample size n is 40.

```
# Set seed to ensure reproduceability
set.seed(123)

# Set lambda to 0.2
lambda <- 0.2

# Generate the sample mean of one simulation with lambda=0.2 and n=40
expo <- rexp(40, lambda)
mean(expo)
```

```
## [1] 4.811212
```

```
# Generate the sample means of 1000 simulations with lambda=0.2 and n=40
expo1000 <- as.data.frame(replicate(1000,mean(rexp(40, lambda))))
names(expo1000) <- c("sample.mean")

# Calculate the mean of this simulation of sample means
```

```r
mean1000 <- mean(expo1000$sample.mean)
mean1000
```

```
## [1] 5.013543
```

```r
# The theoretical mean which is the centre of the distribution is
1/lambda
```

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

The center of distribution of sample means of 40 exponentials is close to the theoretical center of the distribution.

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

```r
# Calculate the variance of this simulation of sample means using theta^2/n
var1000 <- var(expo1000$sample.mean)
var1000
```

```
## [1] 0.6024988
```

```r
# The theoretical variance of the distribution is
((1/lambda)^2)/40
```

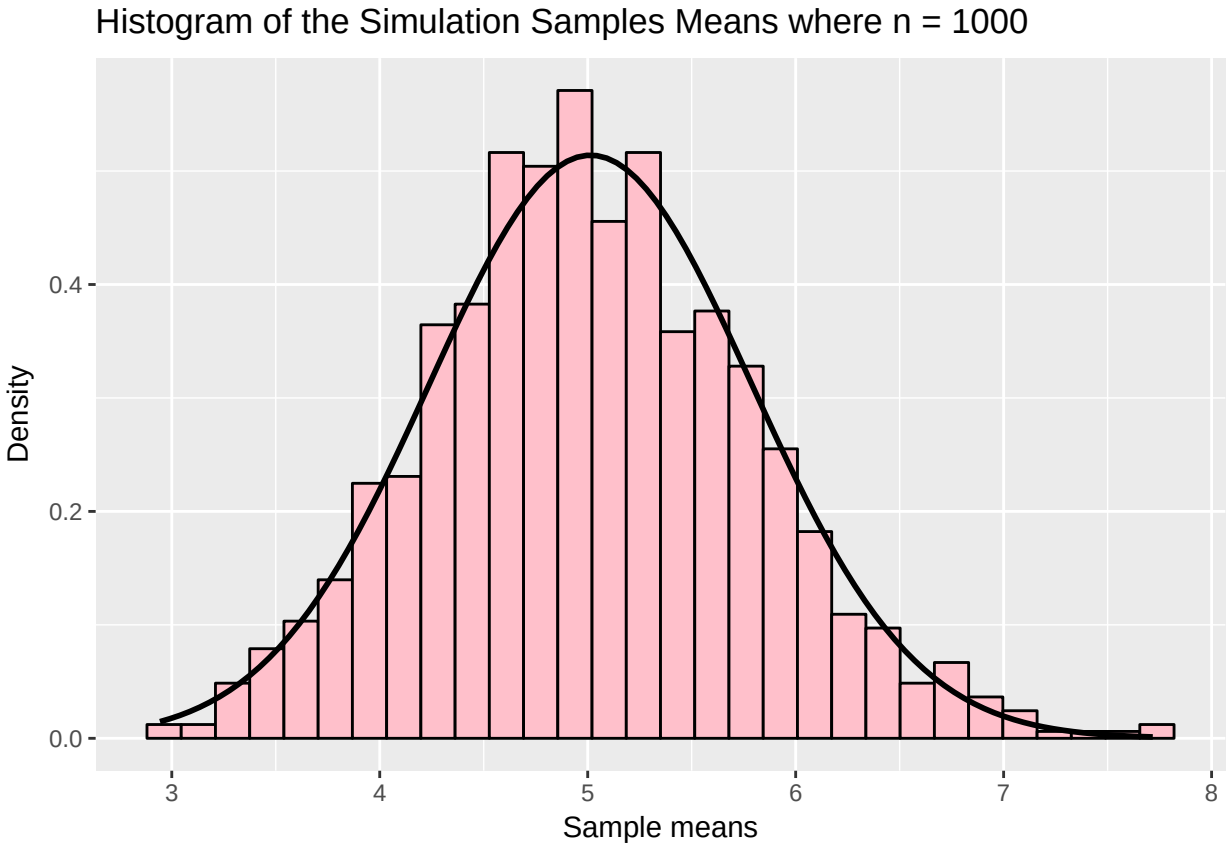
```
## [1] 0.625
```

The variance of the distribution of means of 40 exponentials is less than the theoretical variance of the distribution.

1.3 Show that the distribution is approximately normal.

```r
# Plot the sample means from the simulation.
ggplot(data = expo1000, aes(x=sample.mean)) +
    geom_histogram(aes(y = ..density..),colour="black",fill="pink")+
    stat_function(fun=dnorm,args=list( mean=mean1000, sd=sqrt(var1000)),geom="line",color = "black", si
    ggtitle("Histogram of the Simulation Samples Means where n = 1000") +
    scale_x_continuous("Sample means")+
    ylab("Density")
```

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



A normal distribution follows a bell shaped curve, where the black line represents the calculated Normal Distribution, which we can compare shape of the histogram.

The central limit theorom states that the sample means would become that of a standard normal distribution as the sample size increases whilst meeting the two conditions of independence ($n < 10\%$) and normal, or if skewed distribution, that $n > 30$.
