Statistical Inference Project 1

Balogun Stephen Taiye

November 22, 2015

Contents

Overview

The project seeks to investigate the distribution of averages of 40 exponentials and compare it with the Central Limit Theorem (CLT).This will be with respect to comparing the sample mean and variance with that of the theoretical values given.

Simulations

We are given a sample size (n) to be 40 and a formular $rexp(n, lambda)$ where:
- $rexp$ is R exponential distribution
- n is the sample size
- $lambda(\lambda)$ is the rate for the sample size (rate given as 0.2)

We try to simulate the data to get several 1000 *means* of the data. The formular for the simulation is given as:

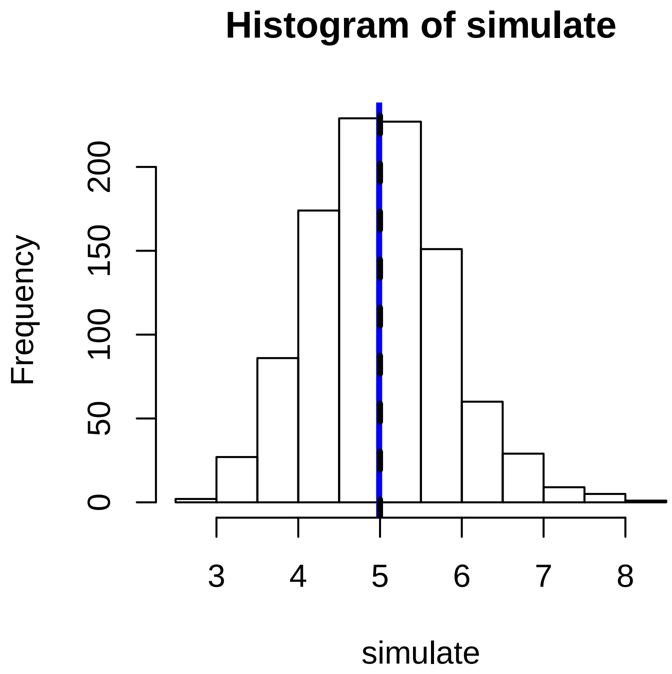
```
nosim <- 1000   ##  number of times to simulate the data
n <- 40   ##   sample size
lambda <- 0.2   ## the given rate
simulate <- apply(matrix(rexp(nosim*n, rate = lambda), nosim), 1, mean)
```

The formular given above samples the given data with replacement and draws sample size of 40 1000 times, then find the mean of those 1000 samples

Sample Mean versus the Theoretical Mean

The $theoreticalmean$ of the R code `rexp(n, lamba)` is given as $1/\lambda$. The formular belows shows a plot that compares this *theoretical mean* with the *mean of our simulated data.*

```
hist(simulate)
abline(v = mean(simulate), lwd = 3, col = "blue")  ## shows the sample mean
abline (v = 1/lambda, lty = "dashed", lwd = 3)   ## shows the theoretical mean on the same plot
```



The solid blue line shows the mean for the simulated data while the black dashed line shows the theoretical mean. The mean of the simulated data only approximates the theoretical mean because the simulation was not done infinite amount of time.

```
simulatedMean <- round(mean(simulate), 3)
theoreticalMean <- round(1/lambda, 3)
cbind(simulatedMean, theoreticalMean)
```

simulatedMean	theoreticalMean
4.989	5

Sample Variance versus Theoretical Variance

The theoretical variance is given as $\sigma^2 = (1/\lambda)^2$ (i.e. $var = sd^2$) which is equal to 25. The variance for our simulated data is given as σ^2/n which can also be expressed as $theoreticalvariance/samplesize(n)$

```
varTheory <- (1/lambda)^2
varSample1 <- round(var(simulate), 3)  ## calculates the variance of our   simulated data
varSample2 <- varTheory / n  ## using the formular (theoretical variance / sample size)
cbind(varSample1, varSample2)  ## compares the two results
```

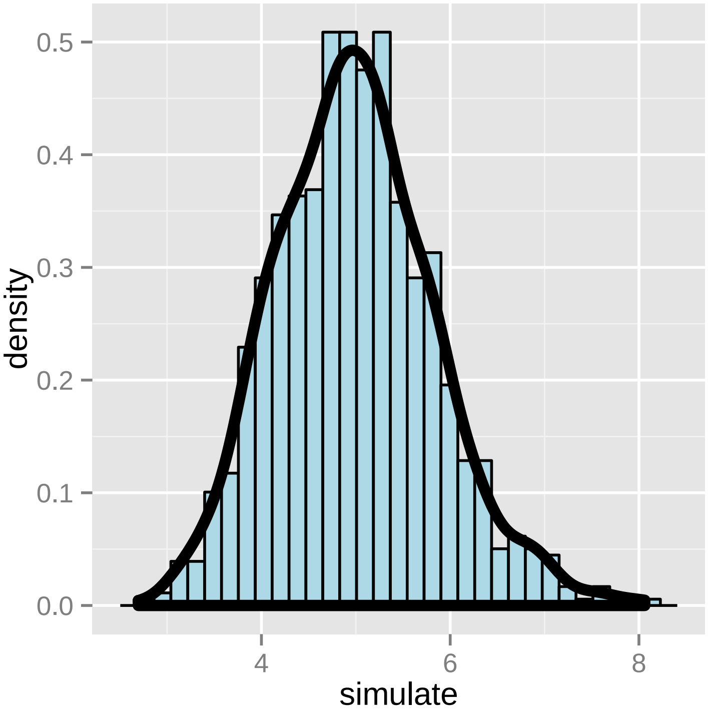
varSample1	varSample2
0.696	0.625

Our calculated variance approximates the expected value when they are rounded up to a decimal place.

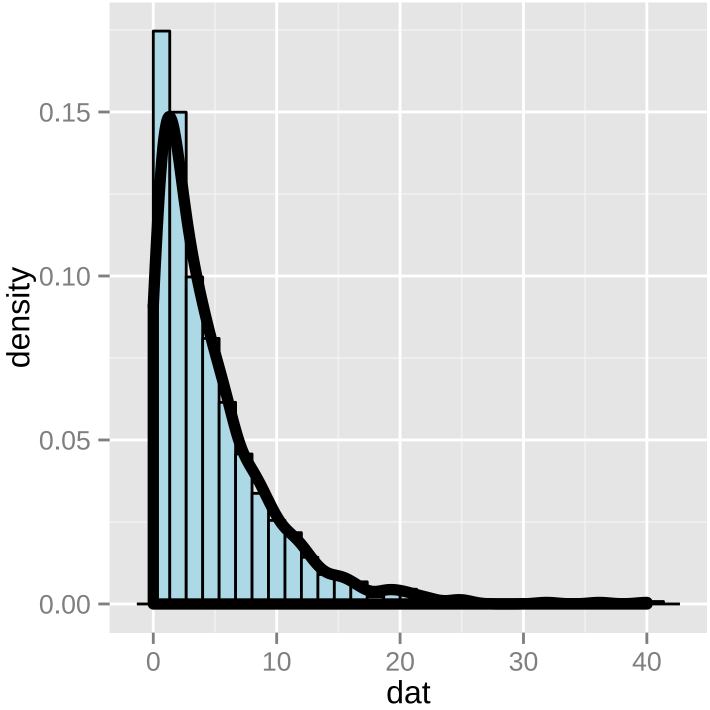
Distribution

This section attempts to compare the distribution of a single sample size (1000) with the distribution of 1000 averages of several sample size (40) and prove that one of them follows approximately a normal distribution. A distribution is said to be approximately normal if it has a Gaussian pattern of distribution (the so-called "bell curve"). The formular below plots a graph of a single sample size (1000) with the averages of sample sizes.

```
##  first i plot the simulated data and overlay the density curve on the histogram
library(ggplot2)
g <- ggplot(data = as.data.frame(simulate), aes(x = simulate))
g <- g + geom_histogram(aes(y = ..density..), fill = "lightblue", colour = "black")
g <- g + geom_density(size = 2, colour = "black")
g
```



```
## next i plot the random exponential of size 1000 and overlay the density curve on the histogram too
dat <- rexp(1000, rate = lambda)   ## generates a random exponential of size 1000
g2 <- ggplot(data = as.data.frame(dat), aes(x = dat))
g2 <- g2 + geom_histogram(aes(y = ..density..), fill = "lightblue", colour = "black")
g2 <- g2 + geom_density(size = 2, colour = "black")
g2
```



From the plots above it becomes obvious that the plot of averages of 40 has a more distribution and hence approximately normal in it's distribution compared to the one of a single sample size of 1000.