Project Report for Coursera Statistical Inference Class

Balogun Stephen Taiye

Sys.Date()

Contents

Project 1

Overview

The project seeks to investigate the distribution of averages of 40 exponentials and compare it with the Central Limit Theorem (CLT)

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 is the rate for the sample size (rate given as 0.2)

Using Bootstrap technique, we try to simulate the data to get several 1000 *means* of the data. The formular for the simulation is given as:

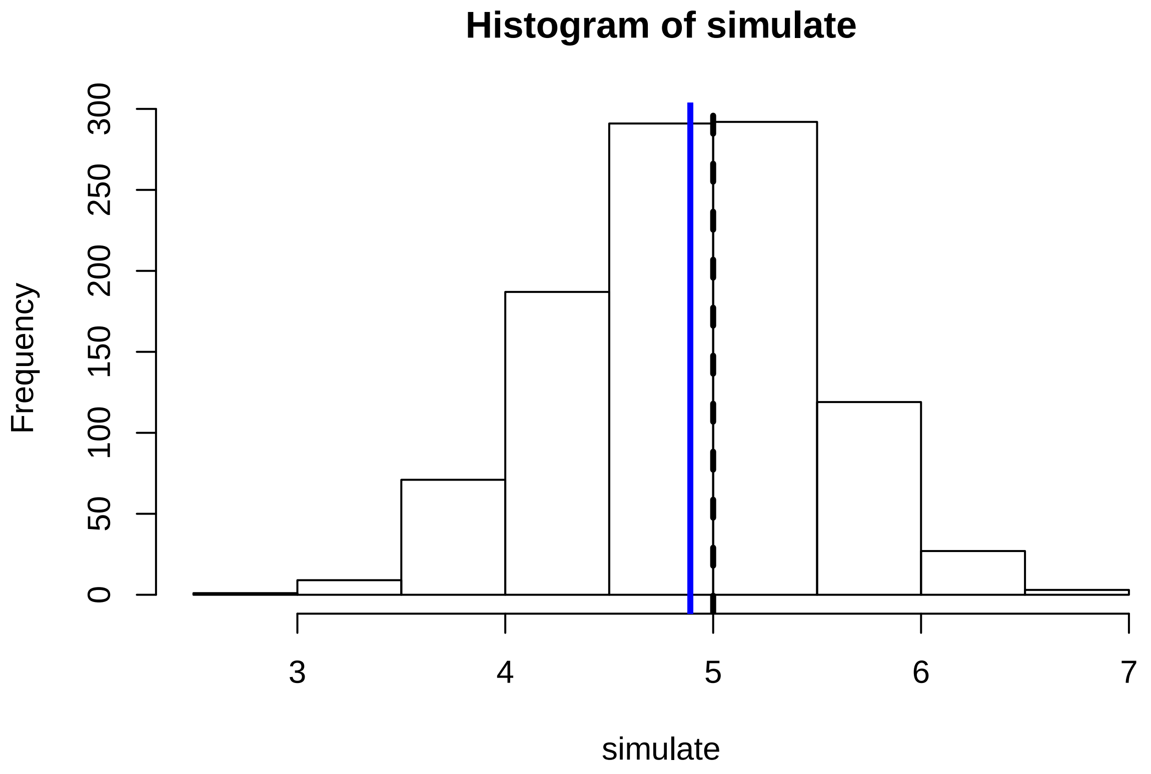
```r
B <- 1000   ##  number of times to simulate the data
n <- 40   ##   sample size
lambda <- 0.2   ## the given rate
simulate <- apply(matrix(sample(rexp(n, lambda), B*n, replace = TRUE),
                         B), 1, mean)
```

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
```



Rounding up the simulated data *mean* 2 decimal places and comparing it with The theoretical mean shows that the simulated mean approximates the theoretical mean

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

```
##      simulatedMean theoreticalMean
## [1,]          4.89               5
```

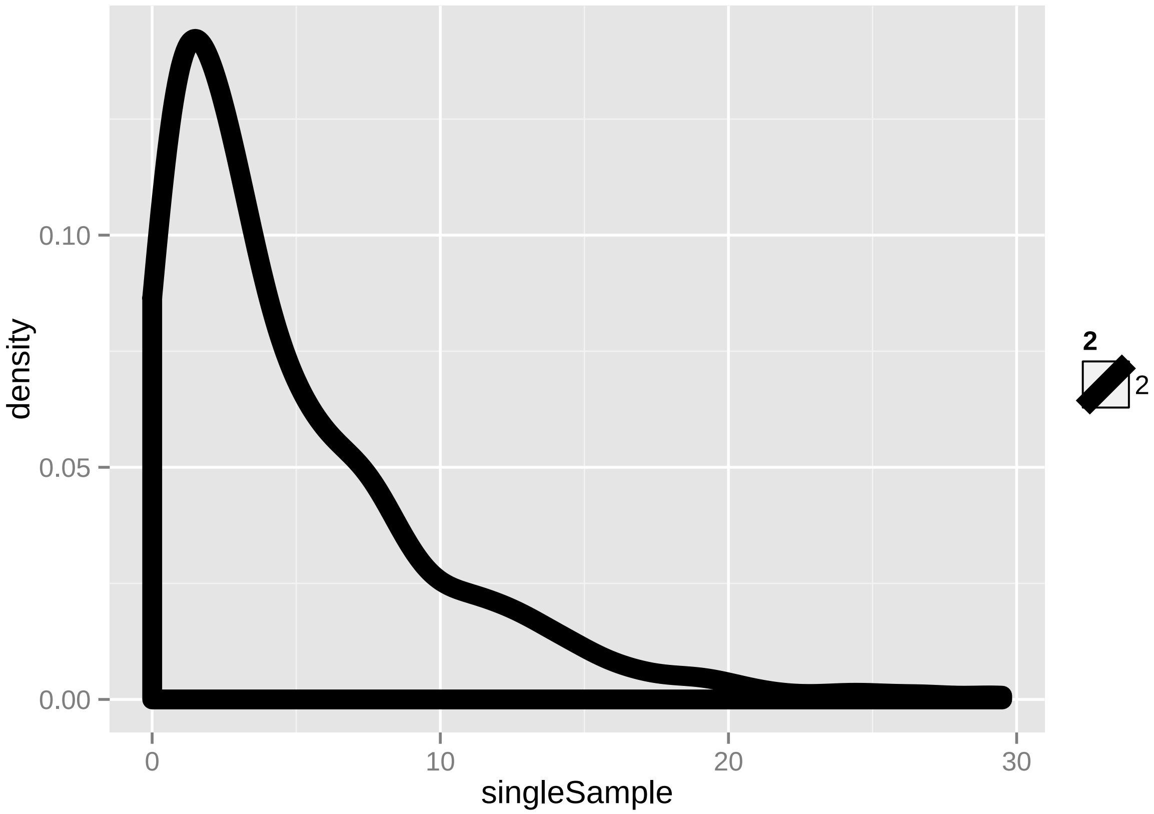
Sample Variance versus Theoretical Variance

The theoretical variance is given as $\sigma = 1/\lambda$ (same formular as the theoretical mean). The formular for calculating the variance for the simulated data is given as $var = \sigma/\sqrt(n)$

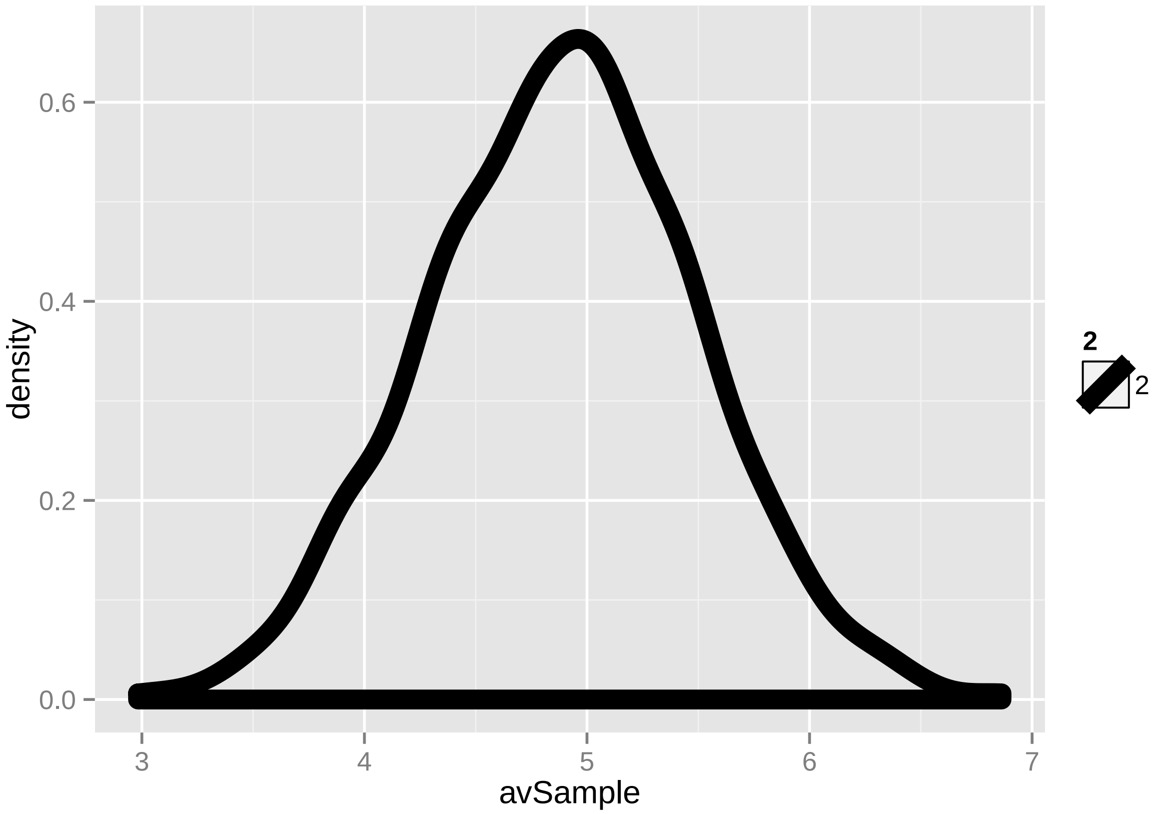
Distribution

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.

```
singleSample <- rexp(1000, 0.2)
avSample <- simulate   ## remember our "simulate" is an average of 1000 samples each of size 40
par(mfrow = c(2, 1))  ## allows side by size plotting
library(ggplot2)
qplot(singleSample, geom = "density", color = I("black"), size = 2)
```



```
qplot(avSample, geom = "density", color = I("black"), size = 2)
```



Project 2

Overview

This second project aims to analyse the *Toothgrowth* data in `R datasets package`, perform some basic exploratory analysis on it, then run some statistical inference with respect to calculation of confidence interval/hypothesis testing

```
library(printr)
Toothdata <- ToothGrowth  ## loads the "toothgrowth" data in the datasets package
dim(Toothdata)    ## shows the number of rows and columns of the data
```

```
## [1] 60  3
```

```
summary(Toothdata)  ##  summarises the data
```

len	supp	dose
Min. : 4.20	OJ:30	Min. :0.500
1st Qu.:13.07	VC:30	1st Qu.:0.500
Median :19.25	NA	Median :1.000
Mean :18.81	NA	Mean :1.167
3rd Qu.:25.27	NA	3rd Qu.:2.000
Max. :33.90	NA	Max. :2.000

```
any(is.na(Toothdata))  ## checks if there are any uncompleted records in the data
```

```
## [1] FALSE
```

```
head(Toothdata)    ## checks the first few rows to give an idea how the data looks like
```

len	supp	dose
4.2	VC	0.5
11.5	VC	0.5
7.3	VC	0.5
5.8	VC	0.5
6.4	VC	0.5
10.0	VC	0.5

```
tail(Toothdata)   ## checks the last few rows to be sure that there are same number of entries for each
```

	len	supp	dose
55	24.8	OJ	2
56	30.9	OJ	2
57	26.4	OJ	2
58	27.3	OJ	2
59	29.4	OJ	2
60	23.0	OJ	2

Now that basic exploration has been done, we try to format the data to allow for our hypothesis testing. The data is divided into three different entries each for each of the `dose` of the supplements.

```
library(dplyr)   ## seems to be my favourite for getting and cleaning data
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
sub0.5 <- filter(Toothdata, dose == .5)  ## creates a subset of dose 0.5
sub1 <- filter(Toothdata, dose == 1.0)   ## creates a subset of dose 1.0
sub2 <- filter(Toothdata, dose == 2.0)   ## creates the subset with dose 2.0
```

Next, we run `t.test()` to compare the t-test for the two different types of supplements at the doses given.

```r
t0.5 <- t.test(len~supp, data = sub0.5)$conf.int   ## t.test at dose 0.5
t1.0 <- t.test(len~supp, data = sub1)$conf.int  ## t.test at dose 1.0
t2.0 <- t.test(len~supp, data = sub2)$conf.int  ## t.test at dose 2.0
rbind(t0.5, t1.0, t2.0)  ## for direct comparison of the different t.test() for the three doses
```

t0.5	1.719057	8.780943
t1.0	2.802148	9.057852
t2.0	-3.798071	3.638070