Statistical Inference Project- Part 1

Van Wyk Burnett

02 April 2020

Part 1: Simulation Exercise

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.6.3
```

```
library(ggplot2)
library(dplyr)
```

```
##
## 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
```

Data Processing

First set the varianes to the given values. Also set the seed to that the similation can be done by another with the same results.

```
set.seed(234)
lambda <- 0.2
n <- 40  #number of exponentials (sample size)
nosim <- 1000 #number of simulatoins
```

Next, create a matrix of 1000 simulations. Each time drawing 40 samples from the exponential distribution.

```
mymatrix <- matrix(rexp(n * nosim, rate = lambda), nosim)
```

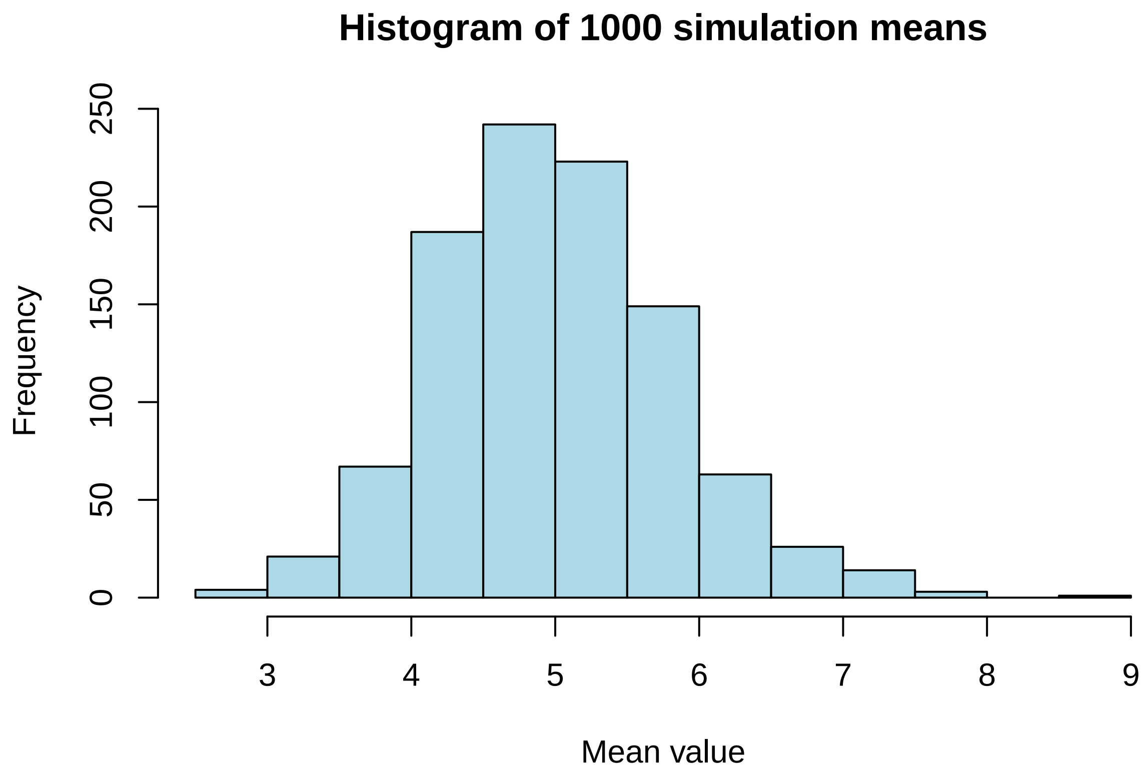
Calculate the mean for each simulation.

```
means <- rowMeans(mymatrix)
```

Exploratory data analysis

Plot the simulation means.

```
hist(means, col="lightblue", xlab = "Mean value", main = "Histogram of 1000 simulation means")
```



Mean Comparison

Question 1: Show the sample mean and compare it to the theoretical mean on the distribution.

```
#mean of sample means
smean <- mean(means)
smean
```

```
## [1] 5.001573
```

```
#theoretical mean
tmean <- 1/lambda
tmean
```

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

```
#difference
smean - tmean
```

```
## [1] 0.00157285
```

The mean of the sample mean is very close to the theoretical mean, with only 0.001573 difference.

Variance Comparison

Question 2: Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.

```
#Vaiance of sample mean
sVar <- var(means)
sVar
```

```
## [1] 0.6631504
```

```
#Theoretical variance
tVar <- (1/lambda)^2/(n)
tVar
```

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

```
#Difference
sVar-tVar
```

```
## [1] 0.03815044
```

The sample variance and the theretical variance is very close to each other, with a difference of only 0.038150.

Standard Deviation

```
#sample standard deviation
sstdev <- sd(means)
sstdev
```

```
## [1] 0.8143405
```

```
#theoretical standard deviation
tstdev <- 1/(lambda * sqrt(n))
tstdev
```

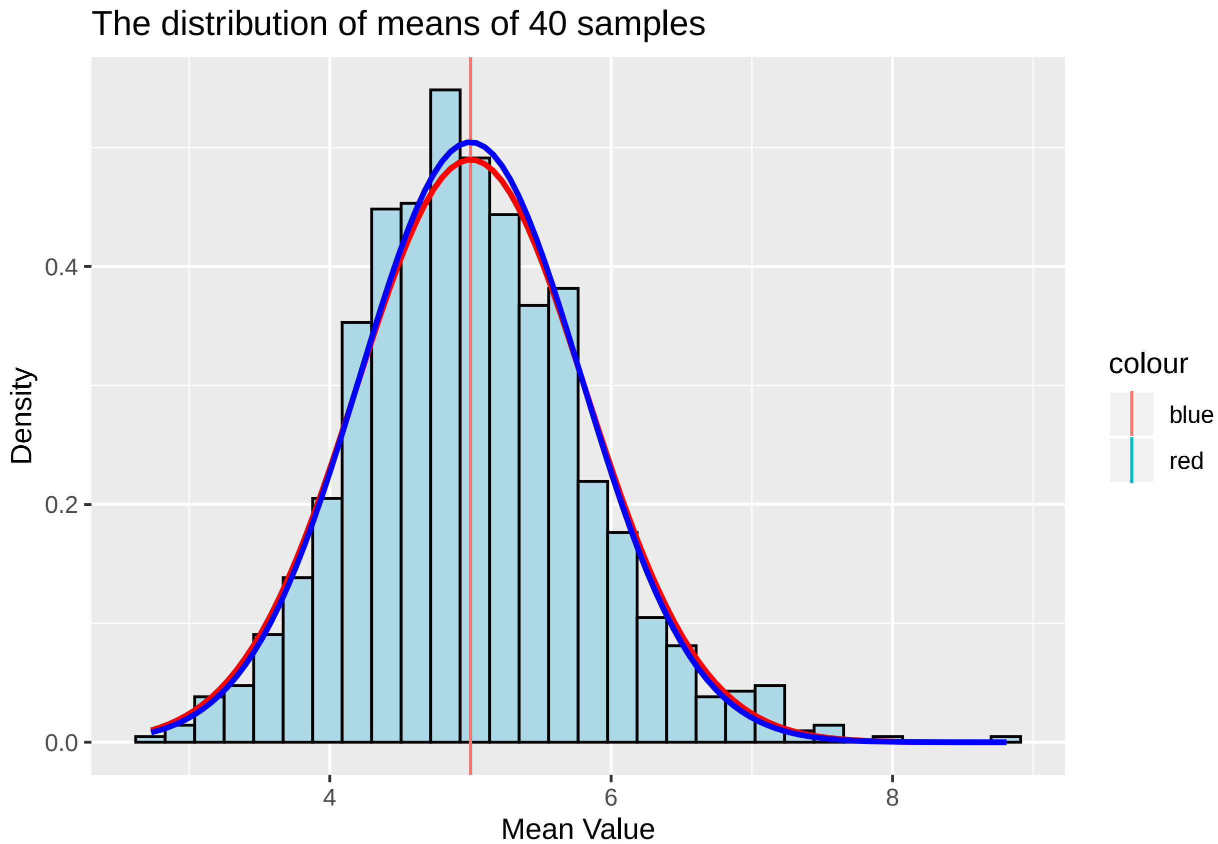
```
## [1] 0.7905694
```

Camparing to the normal distribution

Question 3: Show that the distribution is approximately normal.

```
mydata <- data.frame(means)
p <- ggplot(mydata, aes(x = means))
p <- p + geom_histogram(aes(y=..density..), colour = "black", fill = "lightblue")
p <- p + geom_vline(aes(xintercept = smean, colour="red"))
p <- p + geom_vline(aes(xintercept = tmean, colour="blue"))
p <- p + stat_function(fun = dnorm, args = list(mean = smean, sd = sstdev), color = "red", size = 1.0)
p <- p + stat_function(fun = dnorm, args = list(mean = tmean, sd = tstdev), color = "blue", size = 1.0)
p <- p + labs(title = "The distribution of means of 40 samples", x = "Mean Value", y = "Density")
p
```

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



The theoretical mean and sample mean is so close that the lines overlap, the red vertical line show the mean of means value. The red curve is the cuve formed by the sample mean and sandard deviation. The blue curve represents the normal curve formed by the theoretical mean and standard deviation.

It is clear from this graph that the distribution of means of 40 exponential distributions is very close to the normal distribution.

Confidence Intervals

The sample confidenced interval.

```
sCI <- smean + c(-1,1)*1.96*sstdev/sqrt(n)
sCI
```

```
## [1] 4.749206 5.253940
```

The theoretical confidence interval.

```
tCI <- tmean + c(-1,1)*1.96*tstdev/sqrt(n)
tCI
```

```
## [1] 4.755 5.245
```

The sample and theoretical confidence interval is very close to each other.

Conclusion

It is clear from the evidence shown in the above analysis that the distribution of means of 40 exponential distributions is approximately normal.