Statistical Inference Course Project Parts 1 and 2

Ignacio Ojea

June 2, 2018

Course Project - Part 1

Assignment 1

In this project you will 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.

We begin by assigning the relevant parameters:

```
lambda <- 0.2     # labmda
n.exp <- 40       # number of exponentials
n.sims <- 1000    # number of simulations
set.seed(2018)
```

We then proceed to run the simulations:

```
simulation.data <- as.data.frame(replicate(n.sims, rexp(n.exp, lambda)))
```

We obtained a data frame with 40 rows and 1000 columns; each row corresponding to a simulation.

Assignment 2: Sample Mean vs Theoretical Mean

Let us find the sample means:

```
data.means <- apply(simulation.data, 2, mean)
```

The theoretical mean of exponential distribution is 1/lambda. In our case, this corresponds to $1/(2/10) = 10/2 = 5$. The simulation sample mean is:

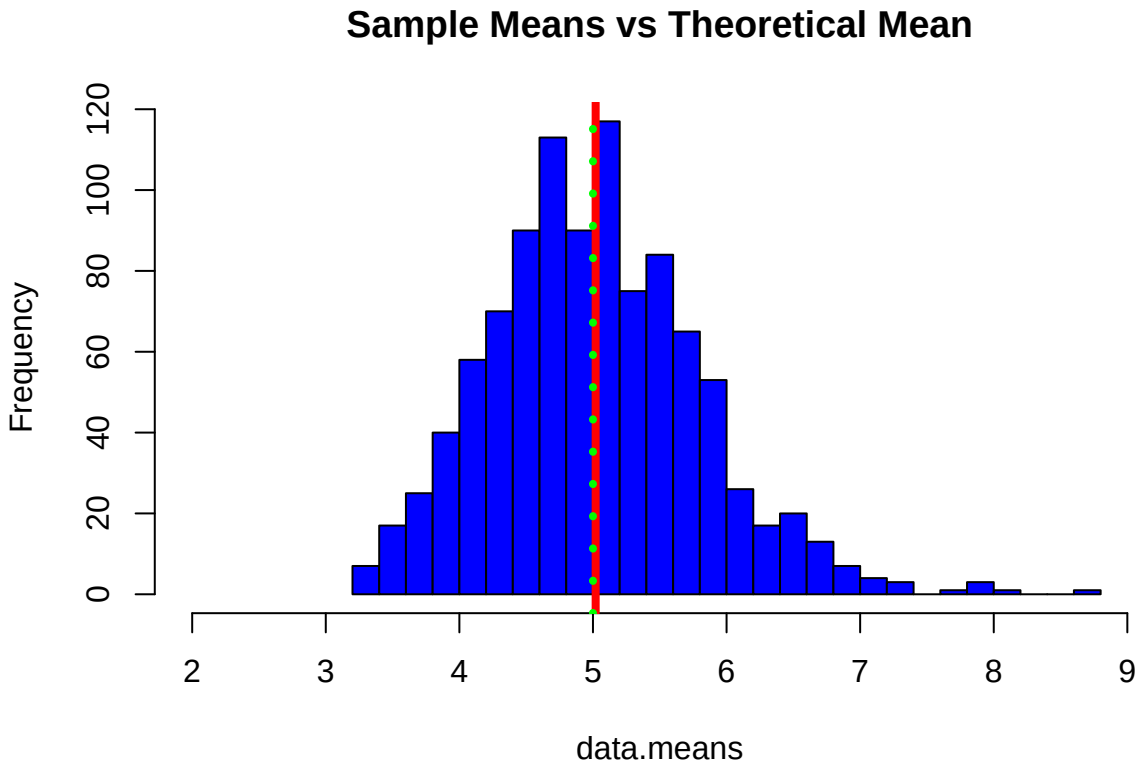
```
mean(data.means)
```

```
## [1] 5.020107
```

Which is very close to the theoretical mean.

Let us now create a plot in order to compare the theoretical mean and the simulations mean:

```
hist(data.means, breaks=30, xlim = c(2,9), main="Sample Means vs Theoretical Mean", col = "blue")
abline(v=mean(data.means), lwd="4", col="red")        # Simulation Mean
abline(v=5, lwd="4", col="green",lty="dotted")        # Theoretical Mean
```



Assignment 3: Sample Variance vs Theoretical Variance From CLT, we know the theoretical standard deviation of the mean is (1 / lambda) / sqrt(n), and the variance is (1 / lambda)^2 / n. Let us now compare:

```
print(paste("Theoretical variance: ", round( ((1/lambda)/sqrt(n.exp))^2 ,5)))
```

```
## [1] "Theoretical variance:  0.625"
```

```
print(paste("Sample variance: ", round(var(data.means) ,5) ))
```

```
## [1] "Sample variance:  0.62613"
```

```
print(paste("Theoretical standard deviation: ", round( (1/lambda)/sqrt(n.exp) ,5)))
```

```
## [1] "Theoretical standard deviation:  0.79057"
```

```
print(paste("Sample standard deviation: ", round( sd(data.means),5)))
```

```
## [1] "Sample standard deviation:  0.79129"
```

The results show that variances are very close, as well as standar deviations.

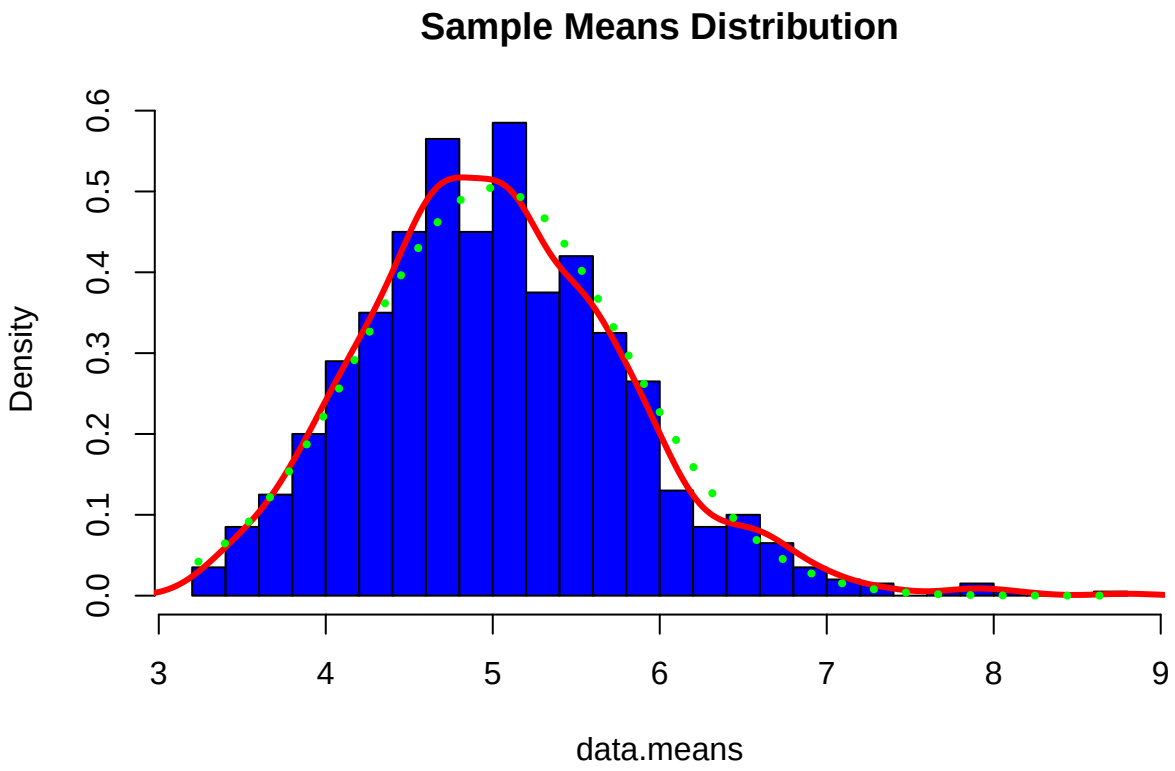
Assignment 2: Approximation to normality

Finally, let us investigate if the exponential distribution is approximately normal. We know from the CLT that the averages of samples should follow a normal distribution as samples increase in number.

```
#Plotting of the mean distribution of the samples
hist(data.means, breaks = 30, prob=TRUE, col="blue", main="Sample Means Distribution")
lines(density(data.means), lwd=3, col="red")

#Plotting of the normal distribution line
x <- seq(min(data.means), max(data.means), length=2*n.exp)
```

```
y <- dnorm(x, mean=1/lambda, sd=sqrt(((1/lambda)/sqrt(n.exp))^2))
lines(x, y, pch=22, col="green", lwd=4, lty = "dotted")
```



The graph shows that the distribution of the sample means (of our exponential distributions) resembles a normal distribution. Once again, this is secured by the Central Limit Theorem, that states that if we were to increase our number of samples (currently 1000), the distribution would be even closer to the standard normal distribution. The green dotted line above is a normal distribution curve and we can see that it is very close to our sampled curve, which is the red line above.

Course Project - Part 2

1. Load the data, basic exploratory analysis, and some formatting

```
set.seed(2018)
data("ToothGrowth")       # Load Data
str(ToothGrowth)           # Structure of data
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
```

2 Provide some basic summary (and visualization)

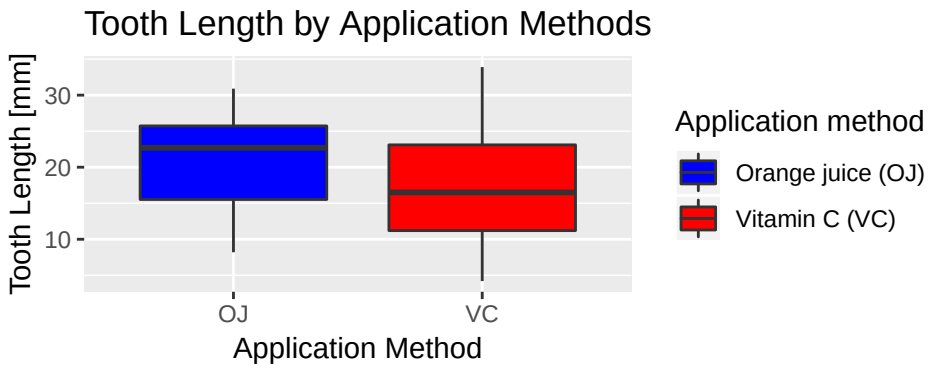
```
summary(ToothGrowth)           # Structure of data
```

```
##       len        supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
##  Median :19.25           2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

```
library(ggplot2)              # In order to provide visualization
mean.supp <- split(ToothGrowth$len, ToothGrowth$supp) # Means by supp
sapply(mean.supp, mean)
```

```
##       OJ       VC
## 20.66333 16.96333
```

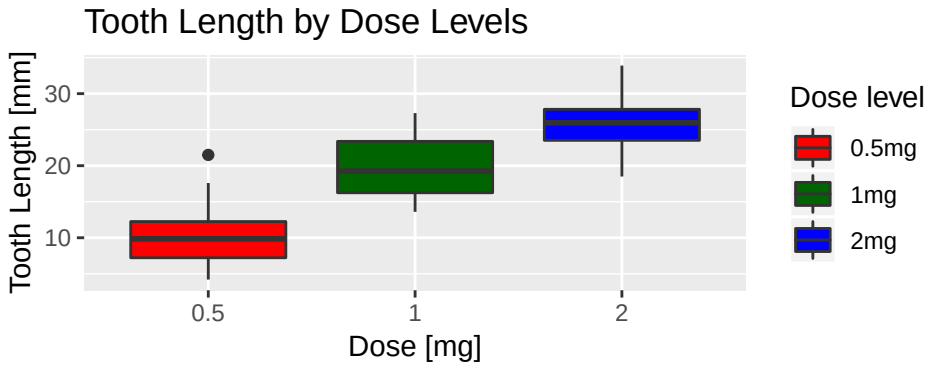
```
ggplot(aes(x = supp, y = len), data = ToothGrowth) + geom_boxplot(aes(fill = supp)) + xlab("Application
```




```
mean.dose <- split(ToothGrowth$len, ToothGrowth$dose) # Means by dose
sapply(mean.dose, mean)
```

```
##    0.5      1      2
## 10.605 19.735 26.100
```

```
ggplot(aes(x = dose, y = len), data = ToothGrowth) + geom_boxplot(aes(fill = dose)) + xlab("Dose [mg]")
```



3 Test to compare tooth growth by supp and dose.

Let us start by comparing tooth growth by supplement using a t-test.

```
t.test(len~supp,data=ToothGrowth)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean in group OJ mean in group VC
##         20.66333         16.96333
```

Since the p-value = 0.06 > 0.05 and the confidence interval [-0.17,7.57] contains zero, we can say that supplement types seems to have no impact on Tooth growth based on this test.

Now, in order to compare tooth growth by dose, we need to look at the different pairs of dose values.

```
# t-test using dose amounts 0.5 and 1.0 [a 0.5 increase in dosage]
ToothGrowth.sub <- subset(ToothGrowth, ToothGrowth$dose %in% c(1.0,0.5))
t.test(len~dose,data=ToothGrowth.sub)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -6.4766, df = 37.986, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.983781  -6.276219
## sample estimates:
## mean in group 0.5   mean in group 1
##            10.605            19.735
```

```
# t-test using dose amounts 1.0 and 2.0 [a 1 increase in dosage]
ToothGrowth.sub <- subset(ToothGrowth, ToothGrowth$dose %in% c(2.0,1.0))
t.test(len~dose,data=ToothGrowth.sub)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -4.9005, df = 37.101, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996481 -3.733519
## sample estimates:
## mean in group 1 mean in group 2
##          19.735          26.100
```

```
# t-test using dose amounts 0.5 and 2.0 [a 1.5 increase in dosage]
ToothGrowth.sub <- subset(ToothGrowth, ToothGrowth$dose %in% c(2.0,0.5))
t.test(len~dose,data=ToothGrowth.sub)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15617 -12.83383
## sample estimates:
## mean in group 0.5   mean in group 2
##            10.605            26.100
```

In all three cases:

1. The p-value was approximately zero, and < 0.05.
2. The confidence interval does not include zero.
3. Furthermore, the higher the increase in dosage, the smaller the p-value of the test.

Based on this result we can assume that the average tooth length increases with an inceasing dose, and therefore the three null hypothesis can be rejected.

Conclusion

Given the following assumptions:

1. The sample is representative of the population.
2. The distribution of the sample means follows the Central Limit Theorem.

By observing the t-test analysis above, we can conclude that: (a) supplement delivery method has no effect on tooth growth/length, but (b) increased dosages do result in increased tooth length.