# Final Project week 4 Statistical Inference

*Daniele Franco de Toledo*

*22 december 2018*

Part 1: Simulation Exercise This part is going to execute simulations and data analysises to illustrate of the central limit theorem.

```
knitr::opts_chunk$set(echo = TRUE)
```

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 also 1/lambda.

```
set.seed(1)
lambda <- 0.2 # Set lambda = 0.2 for all of the simulations.
n <- 40       # In this simulation, we investigate the distribution of averages
              # of 40 exponentials.
simulations <- 1:1000 # We need to do a thousand or so simulated averages
averages <- sapply(simulations, function(x) { mean(rexp(n, lambda)) })
```

## 1. Show where the distribution is centered at and compare it to the theoretical center of the distribution.

When we calculate sample and theorithical mean, we see that both lie close together.

```
mean(averages)
```

```
## [1] 4.990025
```

```
1/lambda
```

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

## 2. Show how variable it is and compare it to the theoretical variance of the distribution.

From the CLT we know that X^bar approximately follows N(mu, sigma^2/n). We know sigma to be 1/lambda. As such it follows that the theoretical standard deviation is:

```
(1/lambda)/sqrt(40) # Theoretical standard deviation
```

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

```
sd(averages)        # actual standard deviation
```

```
## [1] 0.7817394
```

```
# And the variances
((1/lambda)/sqrt(40))^2
```

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

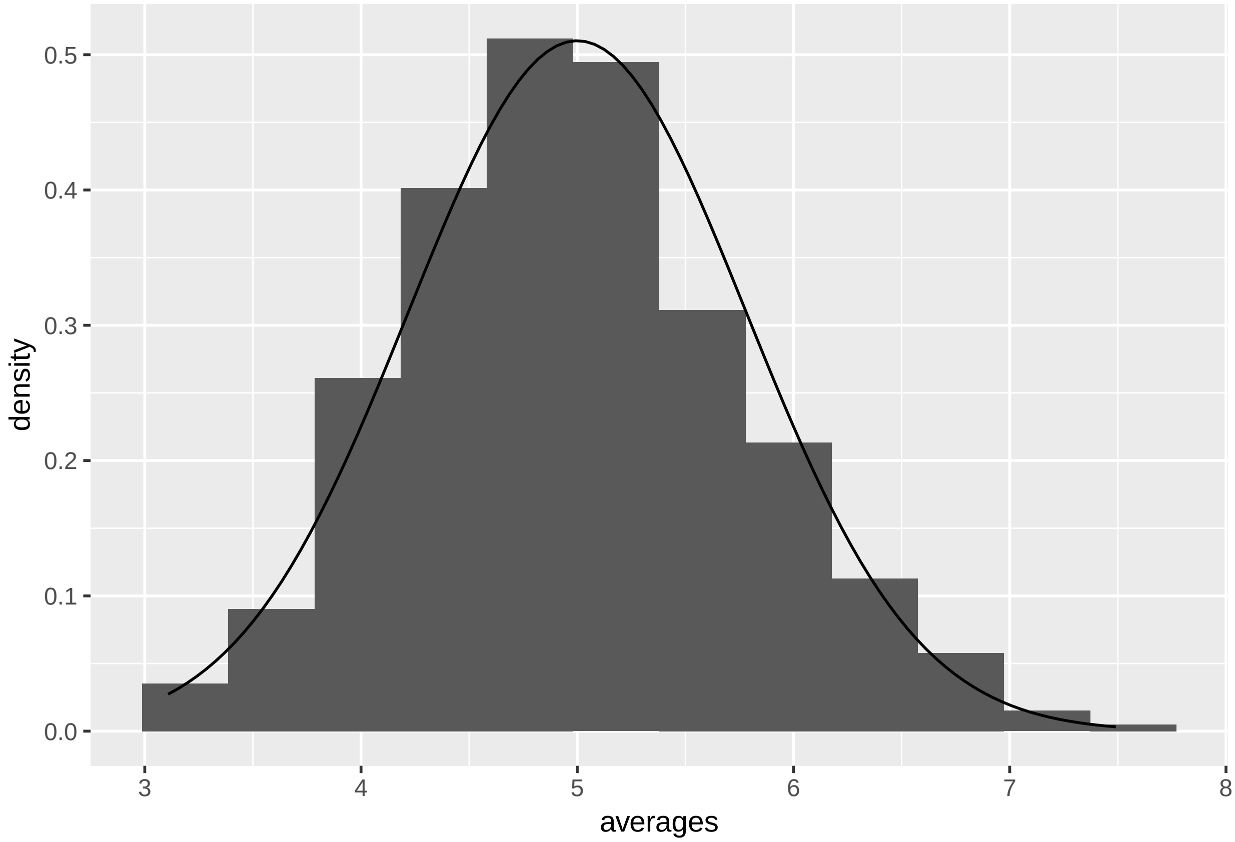
```
sd(averages)^2
```

```
## [1] 0.6111165
```

## 3. Show that the distribution is approximately normal.

To do so, we plot an histogram of thesampled means and overlay the normal distribution with mean 5 and standard deviation 0.7817394 on top of it. We see that the normal distribution indeed closely matches the barplot of the means.

```
library(ggplot2)
# Sturges' formula
k <- ceiling(log2(length(simulations)) + 1)
bw <- (range(averages)[2] - range(averages)[1]) / k
averages.sd <- sd(averages)
p <- ggplot(data.frame(averages), aes(x=averages))
p <- p + geom_histogram(aes(y=..density..), binwidth=bw)
p <- p + stat_function(fun = dnorm, args=list(mean=5, sd=averages.sd))
p
```



## 4. Evaluate the coverage.

Evaluate the coverage of the confidence interval for 1/lambda:

$$\bar{X} \pm 1.96 \frac{S}{\sqrt{n}}$$

.

```
mean(averages) + c(-1,1) * 1.96 * sd(averages) / sqrt(length(averages))
```

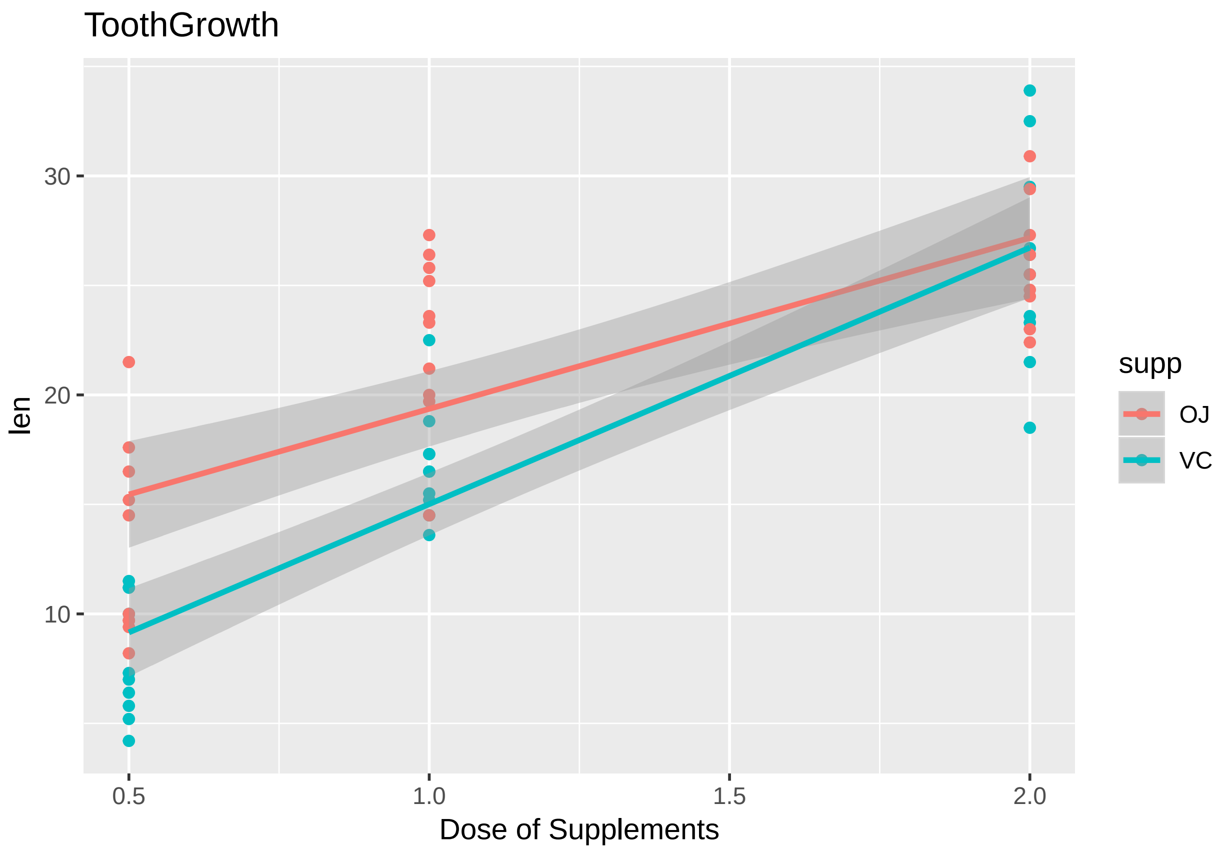
```
## [1] 4.941572 5.038478
```

Part 2: We're going to analyze the ToothGrowth data in the R datasets package.

1. Load the ToothGrowth data and perform some basic exploratory data analyses

We plot the lengt vs the dose for each of the supplements. To gain a better view of groth rates, we also add a loess curve. We see that the growth rates seem to behave differently for both supplements.

```
library(ggplot2)
data(ToothGrowth)
qplot(dose, len, data=ToothGrowth, color = supp, geom = "point") + geom_smooth(method = "lm") + labs(ti
```



2. Provide a basic summary of the data.

This dataset contains three variables: supplement, dose and len. For each supplement, and each dose we calculate basic descriptive statistics: standard deviation, variance, and mean.

```
dose <- as.numeric(levels(as.factor(ToothGrowth$dose)))
supp <- levels(ToothGrowth$supp)
# Structured for further processing
data <- list()
x <- Map(function(s) {
  Map(function(d) {
    l <- ToothGrowth$len[ToothGrowth$dose == d & ToothGrowth$supp == s]
    data <<- rbind(data, list(supp = s, dose = d, sd=sd(l), var=var(l), mu=mean(l)))
  }, dose)
}, supp)
data
```

```
##      supp dose sd       var      mu
## [1,] "OJ" 0.5  4.459709 19.889   13.23
## [2,] "OJ" 1    3.910953 15.29556 22.7
## [3,] "OJ" 2    2.655058 7.049333 26.06
## [4,] "VC" 0.5  2.746634 7.544    7.98
## [5,] "VC" 1    2.515309 6.326778 16.77
## [6,] "VC" 2    4.797731 23.01822 26.14
```

3. Use confidence intervals and hypothesis tests to compare tooth growth by supp and dose. (Use the techniques from class even if there's other approaches worth considering)

We perform the student-t test for each dose level between the two supplements:

```
tests = list()
for (d in dose) {
  ojd <- ToothGrowth$len[ToothGrowth$dose == d & ToothGrowth$supp == "OJ"]
  vcd <- ToothGrowth$len[ToothGrowth$dose == d & ToothGrowth$supp == "VC"]
  t <- t.test(ojd, vcd, var.equal=T)
  id <- paste("OJ", d, "-", "VC", d)
  tests <- rbind(tests, list(id=id, p.value=t$p.value, ci.lo=t$conf.int[1], ci.hi=t$conf.int[2]))
}
tests
```

```
##      id                p.value      ci.lo     ci.hi
## [1,] "OJ 0.5 - VC 0.5" 0.005303661  1.770262  8.729738
## [2,] "OJ 1 - VC 1"     0.0007807262 2.840692  9.019308
## [3,] "OJ 2 - VC 2"     0.9637098    -3.722999 3.562999
```

4. State your conclusions and the assumptions needed for your conclusions.

First, we assume that variance in all groups should be expected to be equal. The underlying assumption is that sampling of Guinea Pigs to assign them to a supplement and a dose was done properly.

Based on the test results from the previous question we need to **reject** the following hypotheses:

- True difference in means between OJ 0.5 and VC 0.5 is equal to 0
- True difference in means between OJ 1 and VC 1 is equal to 0
- True difference in means between OJ 2 and VC 2 is equal to 0