Statistical Inference Course Project

In it, we will use simulation to explore inference and do some simple inferential data analysis. The project consists of two parts:

1. Simulation exercise
2. Basic inferential data analysis

Part 1. Simulation exercise

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. Let's set lambda = 0.2 for all of the simulations. In this simulation, we will investigate the distribution of averages of 40 exponential(0.2)s. Note that we will need to do a thousand or so simulated averages of 40 exponentials.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponential(0.2)s. We will A. Show where the distribution is centered at and compare it to the theoretical center of the distribution. B. Show how variable it is and compare it to the theoretical variance of the distribution. C. Show that the distribution is approximately normal.

```
set.seed(1)
nosim <- 1000
n <- 40
lambda <- .2
sample.mean <- apply(matrix(rexp(nosim * n, lambda), nosim), 1, mean)
```

Distribution properties of sample.mean as follows:

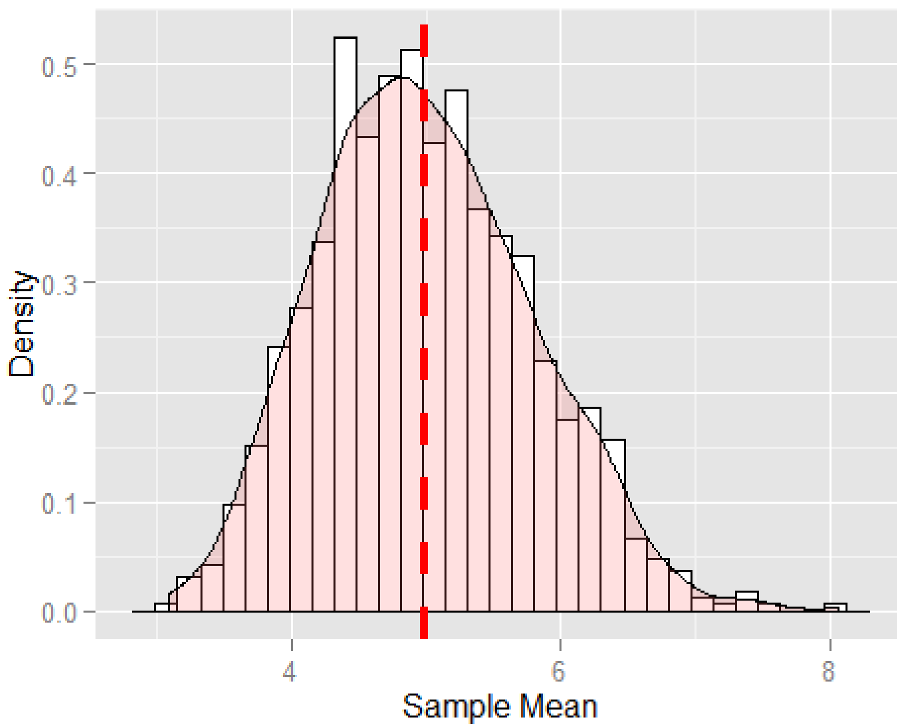
```
range.SM <- range(sample.mean); range.SM

## [1] 3.085 8.059

mean.SM <- mean(sample.mean); mean.SM

## [1] 4.99

sd.SM <- sd(sample.mean); sd.SM

## [1] 0.7859

variance.SM <- sd.SM ^ 2; variance.SM

## [1] 0.6177

median.SM <- qnorm(0.5, mean.SM, sd.SM); median.SM

## [1] 4.99
```

Note that mean and median overlap.

Show where the distribution is centered at and compare it to the theoretical center of the distribution*

We will plot the histogram of the sample means along with the density curve to illustrate the shape and a vertical line indicating the average of sample means.

```
library(ggplot2)
df <- data.frame(x = 1:nosim, y = sample.mean)
ggplot(df, aes(x = sample.mean)) +
        geom_histogram(aes(y = ..density..),
                       binwidth = abs(range.SM[1] - range.SM[2])/30, color =
"black", fill = "white") +
        geom_density(alpha = 0.2, fill = "#FF6666") +
        geom_vline(aes(xintercept = mean.SM), color = "red", linetype =
"dashed", size = 1.5) +
        labs(y = "Density", x = "Sample Mean")
```



For our analysis, the properties of the exponential distribution as follows:

```
mu = 1 / lambda; mu

## [1] 5

sd.E = 1 / lambda; sd.E

## [1] 5
```

```
variance.E = sd.E^2; variance.E

## [1] 25
```

The theory says that this distribution is centered around the population mean (mu).

Part 2. Basic inferential data analysis

Load data.

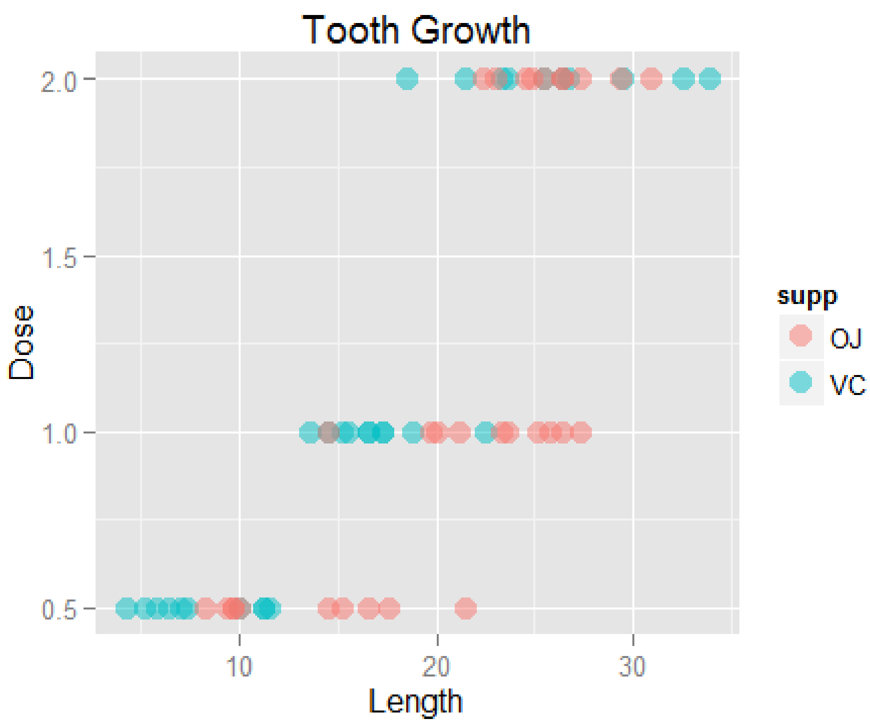
```
library(datasets)
data(ToothGrowth)
```

Let's explore the data.

```
head(ToothGrowth)

##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5

str(ToothGrowth)

## '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 ...

summary(ToothGrowth)

##       len        supp         dose      
##  Min.   : 4.2   OJ:30   Min.   :0.50  
##  1st Qu.:13.1   VC:30   1st Qu.:0.50  
##  Median :19.2           Median :1.00  
##  Mean   :18.8           Mean   :1.17  
##  3rd Qu.:25.3           3rd Qu.:2.00  
##  Max.   :33.9           Max.   :2.00  

table(ToothGrowth$supp, ToothGrowth$dose)

##     
##      0.5  1  2
##   OJ  10 10 10
##   VC  10 10 10
```

We describe that ToothGrowth is a data frame with three variables - length in millimeter, OJ and VC supplements, and doses (0.5, 1,0 and 2.0). Two types of supplements are administered to two groups of 30 subjects each to promote tooth growth. Each supplement is given in 0.5, 1.0 and 2.0 doses to ten subjects each.

```r
library(ggplot2)
ggplot(ToothGrowth, aes(len, dose)) + geom_point(aes(color = supp), size = 4,
alpha = 1/2) + labs(title = "Tooth Growth") + labs(x = "Length", y = "Dose")
```



- It appears that increasing dose (0.5 -> 1.0 -> 2.0) results in longer growth.
- Which supplement is better?
- We will not assumt that higher than 2.0 doses results in even longer growth.

Different doses

Null hypothesis: The means in different doses are the same.

Alternative hypothesis: The means are different.

The groups have the same variance.

Compare 0.5 and 1.0 dosages

```r
TG0.5and1.0 = subset(ToothGrowth, dose %in% c(0.5, 1.0))
t.test(len ~ dose, paired = FALSE, var.equal = TRUE, data = TG0.5and1.0)

##
##  Two Sample t-test
##
## data:  len by dose
## t = -6.477, df = 38, p-value = 1.266e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -11.984  -6.276
## sample estimates:
## mean in group 0.5   mean in group 1
##             10.61             19.73
```

The test suggest that we reject the null hypothesis in favor of the alternative based on p-value = 1.266e-07. The difference in means (-9.13) would only happen by chance 1 time in 1.266 * 10^7 experiments. Thus, we conclude that the difference in means is not a chance, but it is due to a real effect of higher dosage.

Comparing 1.0 and 2.0 dosages yields a similar result.

```
TG1.0and2.0 = subset(ToothGrowth, dose %in% c(1.0, 2.0))
t.test(len ~ dose, paired = FALSE, var.equal = TRUE, data = TG1.0and2.0)
```

```
##
##  Two Sample t-test
##
## data:  len by dose
## t = -4.901, df = 38, p-value = 1.811e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.994 -3.736
## sample estimates:
## mean in group 1 mean in group 2
##           19.73           26.10
```

Which supplement is better?

The two supplements will have differenct variances.

Compare OJ and VC

```
t.test(len ~ supp, paired = FALSE, var.equal = FALSE, data = ToothGrowth)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 1.915, df = 55.31, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.171  7.571
## sample estimates:
## mean in group OJ mean in group VC
##            20.66            16.96
```

OJ is better and the difference ~4 mm lies within the 95% confidence interval.