Simulation and Inferential Data Analysis

PART 1

This is the project for the statistical inference class. In it, you will use simulation to explore inference and do some simple inferential data analysis. The project consists of two parts:

1. A simulation exercise.
2. Basic inferential data analysis.

You will create a report to answer the questions. Given the nature of the series, ideally you'll use knitr to create the reports and convert to a pdf. (I will post a very simple introduction to knitr). However, feel free to use whatever software that you would like to create your pdf.

Each pdf report should be no more than 3 pages with 3 pages of supporting appendix material if needed (code, figures, etcetera).

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.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponential(0.2)s. You should

1. Show the sample mean and compare it to the theoretical mean of the distribution.

```
setwd("~/Desktop/Coursera/StatInf/courseProject")
library(lattice)
library(gplots)
```

```
##
## Attaching package: 'gplots'
##
## The following object is masked from 'package:stats':
##
##     lowess
```

```
lambda <- 0.2
n <- 40
simulations <- 1:2000
set.seed(55) #this is needed to make the exercise reproducible (can be any numb
er)
means <- data.frame(x = sapply(simulations, function(x) {mean(rexp(n, lambd
a))}))
```

```
mean(means$x) # sampling mean = 5.015019
```

```
## [1] 5.015019
```

```
(1/lambda) # theoretical mean = 5
```

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

The distribution is centered at 5.015019 (sampling mean) which is similar to the theoretical mean which is 5

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

```
var(means$x) # sampling var = 0.6212932
```

```
## [1] 0.6212932
```

```
sd(means$x) # sampling sd = 0.7882215
```

```
## [1] 0.7882215
```

```
((1/lambda)/sqrt(n))^2 # theoretical var = 0.625
```

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

```
(1/lambda)/sqrt(n) # theoretical sd = 0.7905694
```

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

Sampling variance and standard deviation are 0.6212932 and 0.7882215 respectively (sampling var and sd)

Theoretical variance and standard deviation are 0.625 and 0.7905694 respectively (theoretical var and sd)

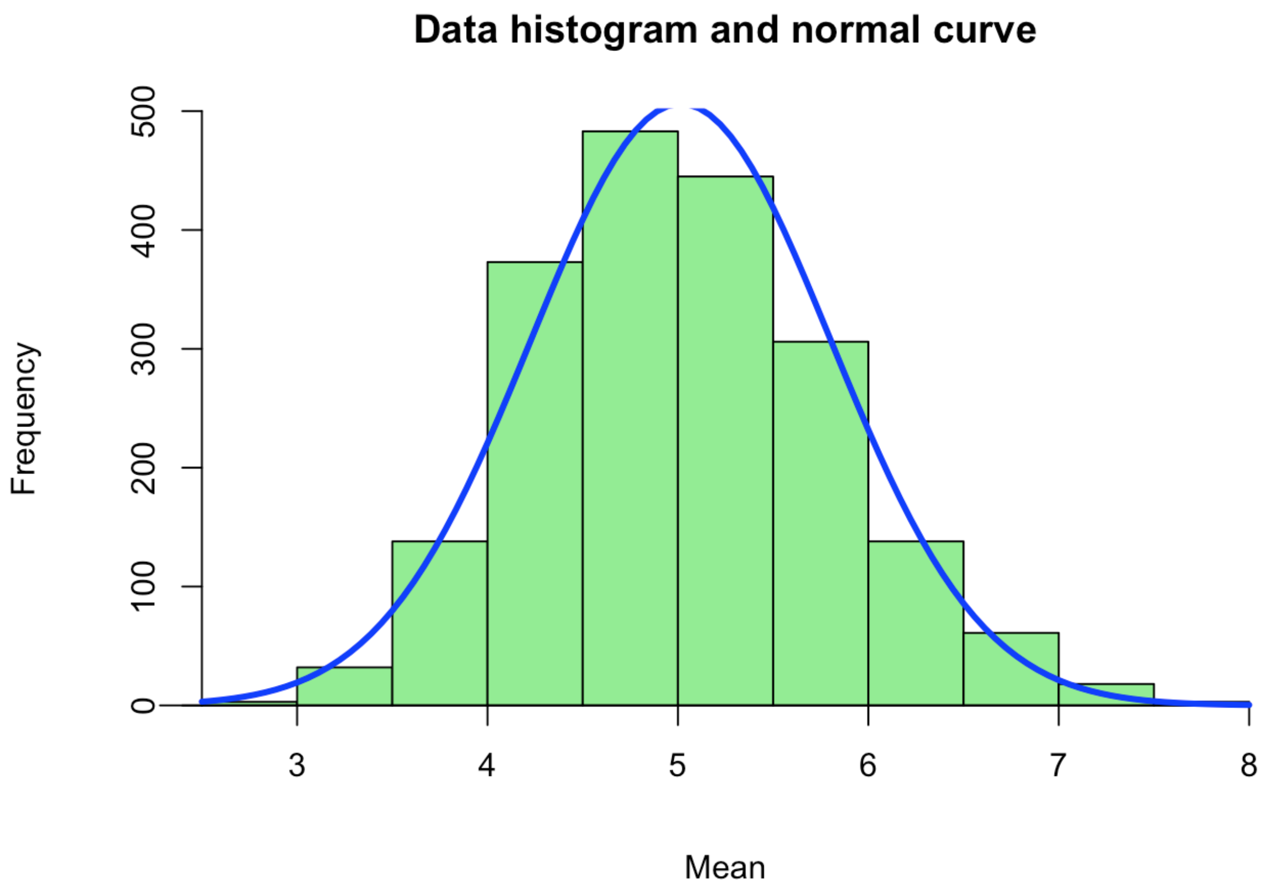
3. Show that the distribution is approximately normal.

A plot is useless, some plot don’t look normal but they are normal and some look normal but they aren’t normal, instead of a plot I’ll use a normality test. After performing a test a plot can give some support to the test.

```
shapiro.test(means$x) # p-value = 2.972e-08
```

```
##
##  Shapiro-Wilk normality test
##
## data:  means$x
## W = 0.9929, p-value = 2.972e-08
```

$p < 0.05$ means the data is represented by a normal curve, which is the case



PART 2

Now in the second portion of the class, we're going to analyze the ToothGrowth data in the R datasets package.

1. Load the ToothGrowth data and perform basic exploratory data analysis

```
data(ToothGrowth)
nrow(ToothGrowth)
```

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

```
ncol(ToothGrowth)
```

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

```
str(ToothGrowth)
```

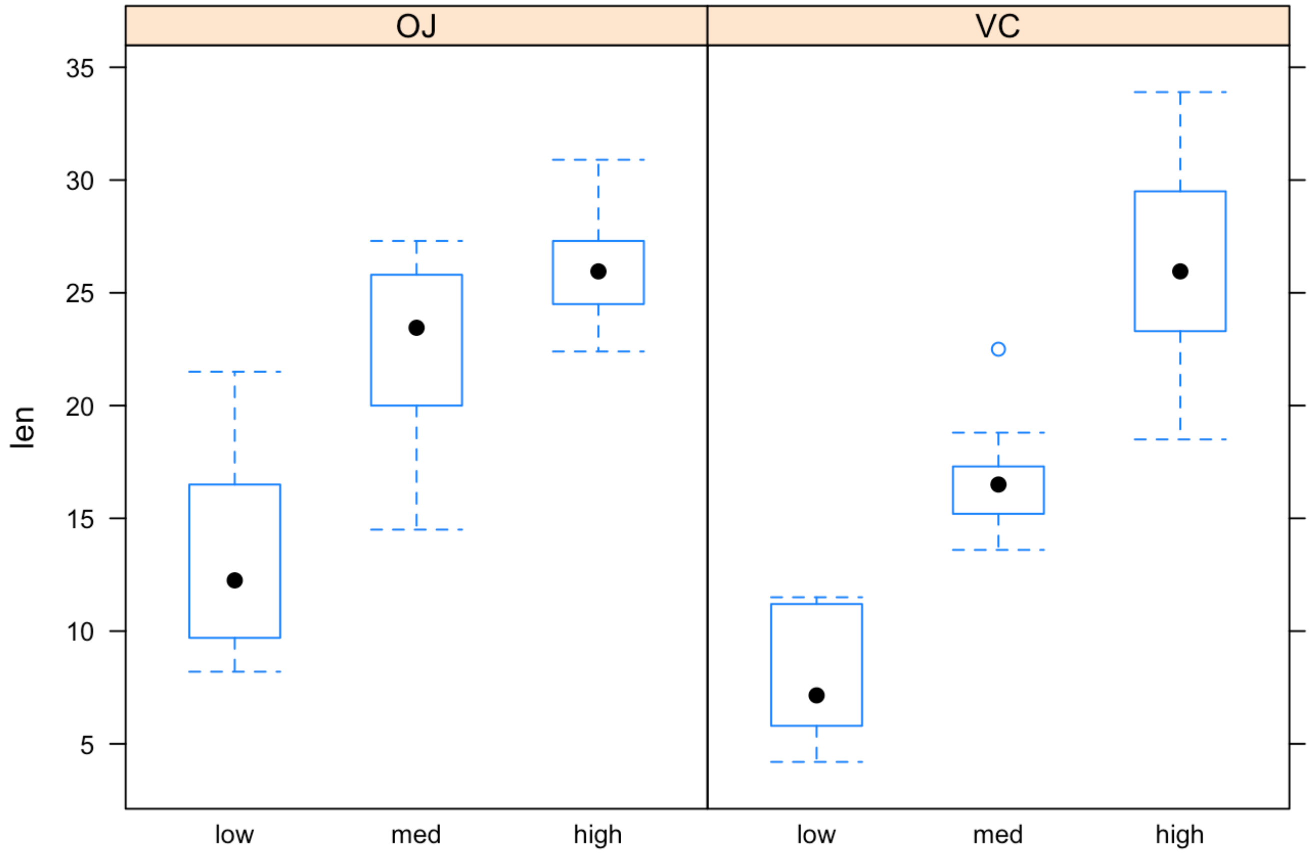
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

```
ToothGrowth$dose
```

```
##  [1] 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 1.0 1.0 1.0 1.0 1.0 1.0 1.0
## [18] 1.0 1.0 1.0 2.0 2.0 2.0 2.0 2.0 2.0 2.0 2.0 2.0 2.0 0.5 0.5 0.5 0.5
## [35] 0.5 0.5 0.5 0.5 0.5 0.5 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 2.0
## [52] 2.0 2.0 2.0 2.0 2.0 2.0 2.0 2.0 2.0
```

```
ToothGrowth$dose = factor(ToothGrowth$dose, levels=c(0.5,1.0,2.0),
                          labels=c("low","med","high"))
attach(ToothGrowth)
table(supp,dose)
```

```
##     dose
## supp low med high
##   OJ  10  10   10
##   VC  10  10   10
```



2. Provide a basic summary of the data.

```
summary(ToothGrowth)
```

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

```
aggregate(len,list(supp,dose), mean)
```

```
##   Group.1 Group.2     x
## 1      OJ     low 13.23
## 2      VC     low  7.98
## 3      OJ     med 22.70
## 4      VC     med 16.77
## 5      OJ    high 26.06
## 6      VC    high 26.14
```

```
aggregate(len,list(supp,dose), sd)
```

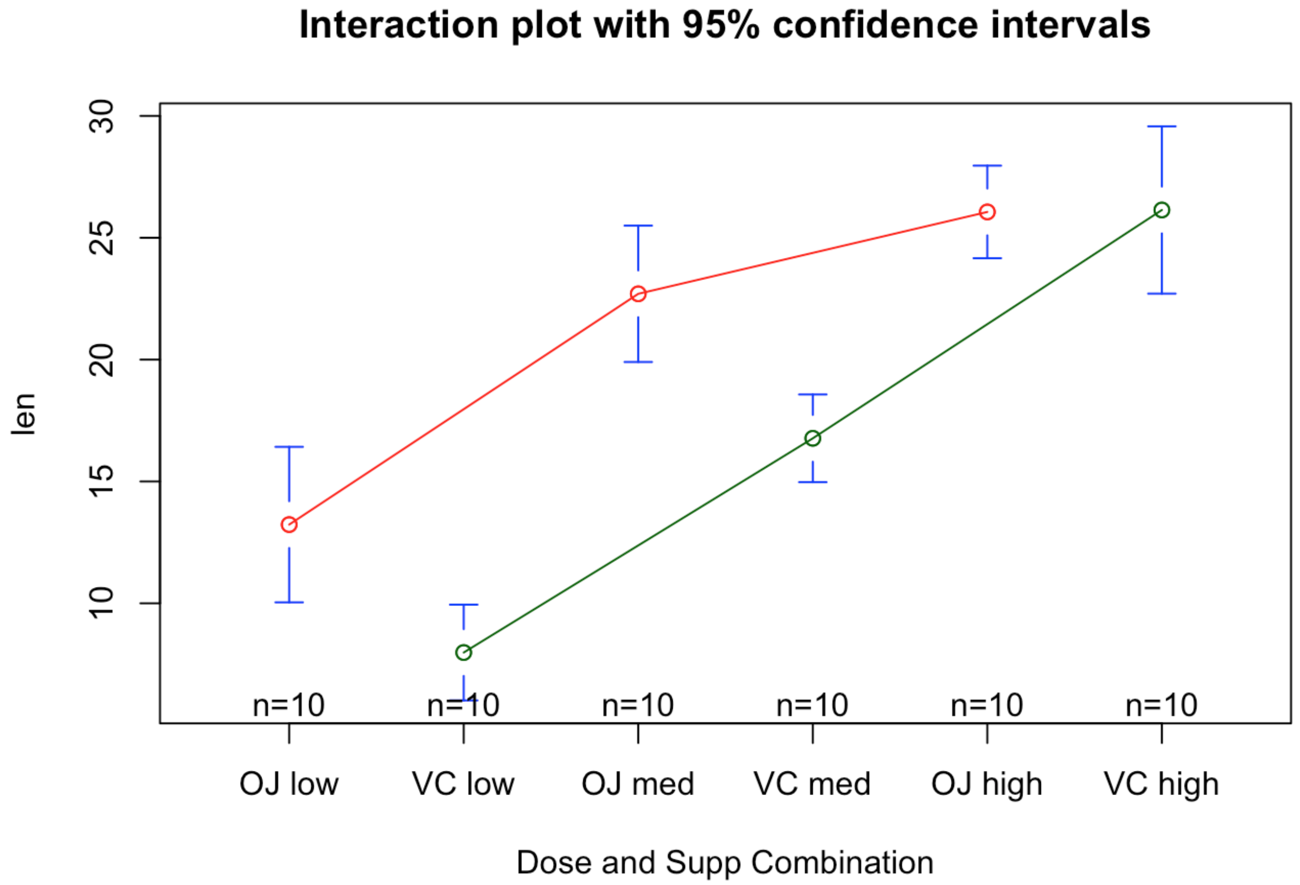
```
##   Group.1 Group.2        x
## 1      OJ     low 4.459709
## 2      VC     low 2.746634
## 3      OJ     med 3.910953
## 4      VC     med 2.515309
## 5      OJ    high 2.655058
## 6      VC    high 4.797731
```

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)

Working with a factor (dose), hence Anova will be useful.

```
anova <- aov(len ~ supp * dose, data=ToothGrowth)
TukeyHSD(anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = len ~ supp * dose, data = ToothGrowth)
##
## $supp
##       diff       lwr       upr     p adj
## VC-OJ -3.7 -5.579828 -1.820172 0.0002312
##
## $dose
##            diff       lwr       upr   p adj
## med-low   9.130  6.362488 11.897512 0.0e+00
## high-low 15.495 12.727488 18.262512 0.0e+00
## high-med  6.365  3.597488  9.132512 2.7e-06
##
## $`supp:dose`
##                  diff        lwr        upr     p adj
## VC:low-OJ:low   -5.25 -10.048124 -0.4518762 0.0242521
## OJ:med-OJ:low    9.47   4.671876 14.2681238 0.0000046
## VC:med-OJ:low    3.54  -1.258124  8.3381238 0.2640208
## OJ:high-OJ:low  12.83   8.031876 17.6281238 0.0000000
## VC:high-OJ:low  12.91   8.111876 17.7081238 0.0000000
## OJ:med-VC:low   14.72   9.921876 19.5181238 0.0000000
## VC:med-VC:low    8.79   3.991876 13.5881238 0.0000210
## OJ:high-VC:low  18.08  13.281876 22.8781238 0.0000000
## VC:high-VC:low  18.16  13.361876 22.9581238 0.0000000
## VC:med-OJ:med   -5.93 -10.728124 -1.1318762 0.0073930
## OJ:high-OJ:med   3.36  -1.438124  8.1581238 0.3187361
## VC:high-OJ:med   3.44  -1.358124  8.2381238 0.2936430
## OJ:high-VC:med   9.29   4.491876 14.0881238 0.0000069
## VC:high-VC:med   9.37   4.571876 14.1681238 0.0000058
## VC:high-OJ:high  0.08  -4.718124  4.8781238 1.0000000
```



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

Assumptions: The sample components are independent and are identically distributed.

Conclussion: Tukey contrast reveals there is no statistical evidence for other ‘noisy’ variables which are relevant, therefore changes in dose and size (independent variables) won’t affect length (dependent variable)