Coursera_DS_Inference_Project2

Atul

January 31, 2016

LOAD THE ToothGrowth DataBase

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

BASIC INFORMATION ABOUT THIS DATABASE

```
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 ...
```

```
colnames(ToothGrowth)
```

```
## [1] "len"  "supp" "dose"
```

```
rownames(ToothGrowth)
```

```
##  [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12" "13" "14"
## [15] "15" "16" "17" "18" "19" "20" "21" "22" "23" "24" "25" "26" "27" "28"
## [29] "29" "30" "31" "32" "33" "34" "35" "36" "37" "38" "39" "40" "41" "42"
## [43] "43" "44" "45" "46" "47" "48" "49" "50" "51" "52" "53" "54" "55" "56"
## [57] "57" "58" "59" "60"
```

```
dim(ToothGrowth)
```

```
## [1] 60  3
```

```
summary(ToothGrowth)
```

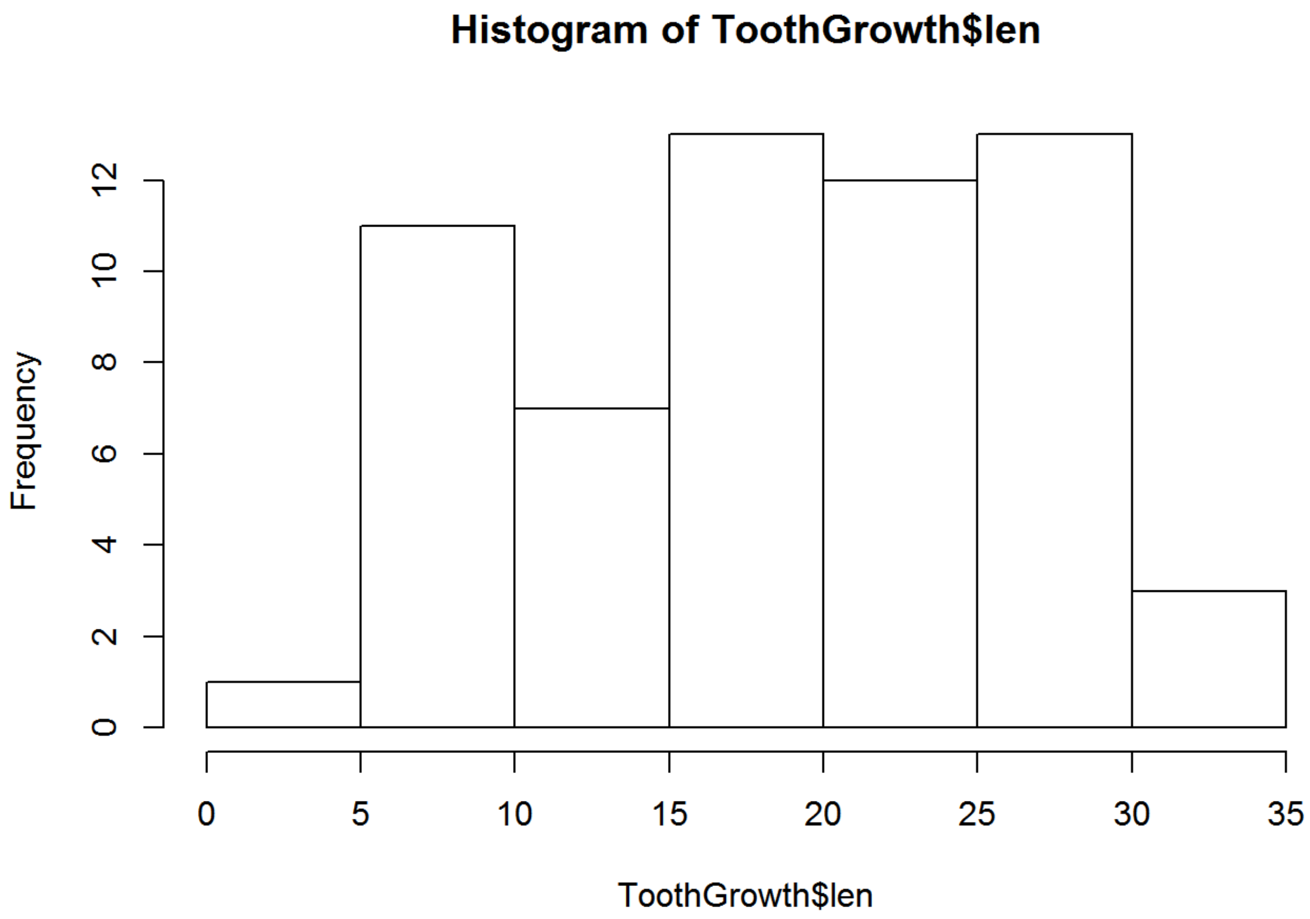
```
##       len        supp         dose      
##  Min.   : 4.20   OJ:30   Min.   :0.500  
##  1st Qu.:13.07   VC:30   1st Qu.:0.500  
##  Median :19.25           Median :1.000  
##  Mean   :18.81           Mean   :1.167  
##  3rd Qu.:25.27           3rd Qu.:2.000  
##  Max.   :33.90           Max.   :2.000
```

```
unique(ToothGrowth$dose)
```

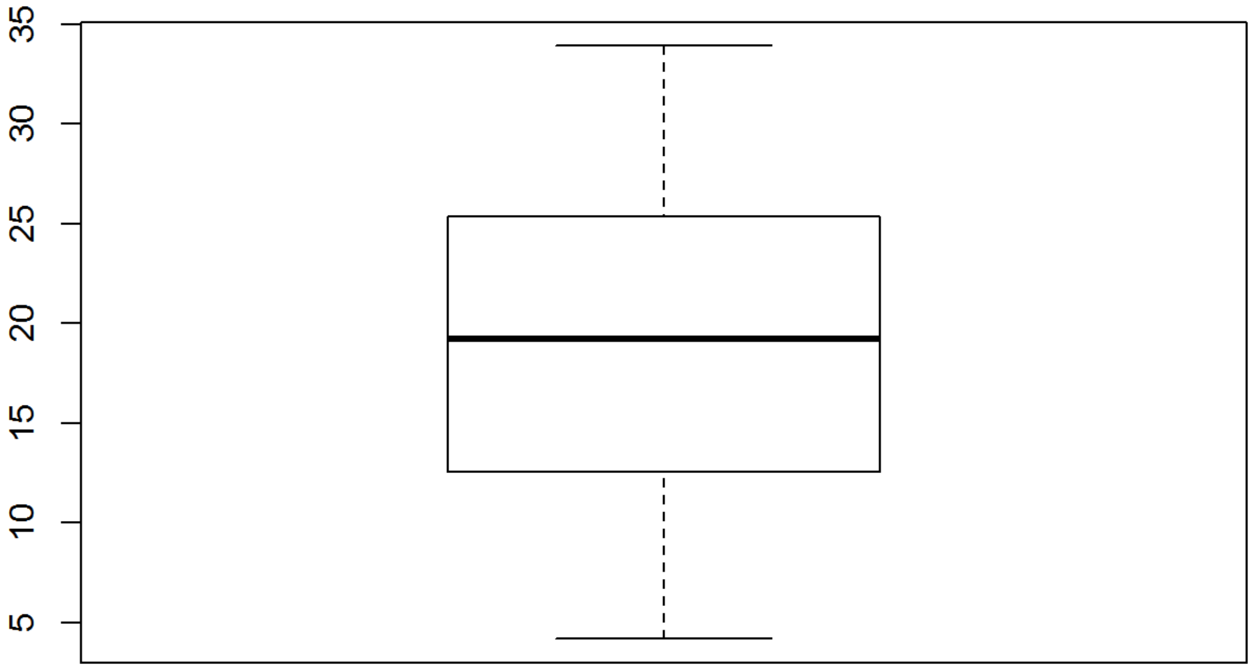
```
## [1] 0.5 1.0 2.0
```

EXPLORATORY DATA ANALYSIS

```
hist(ToothGrowth$len)
```

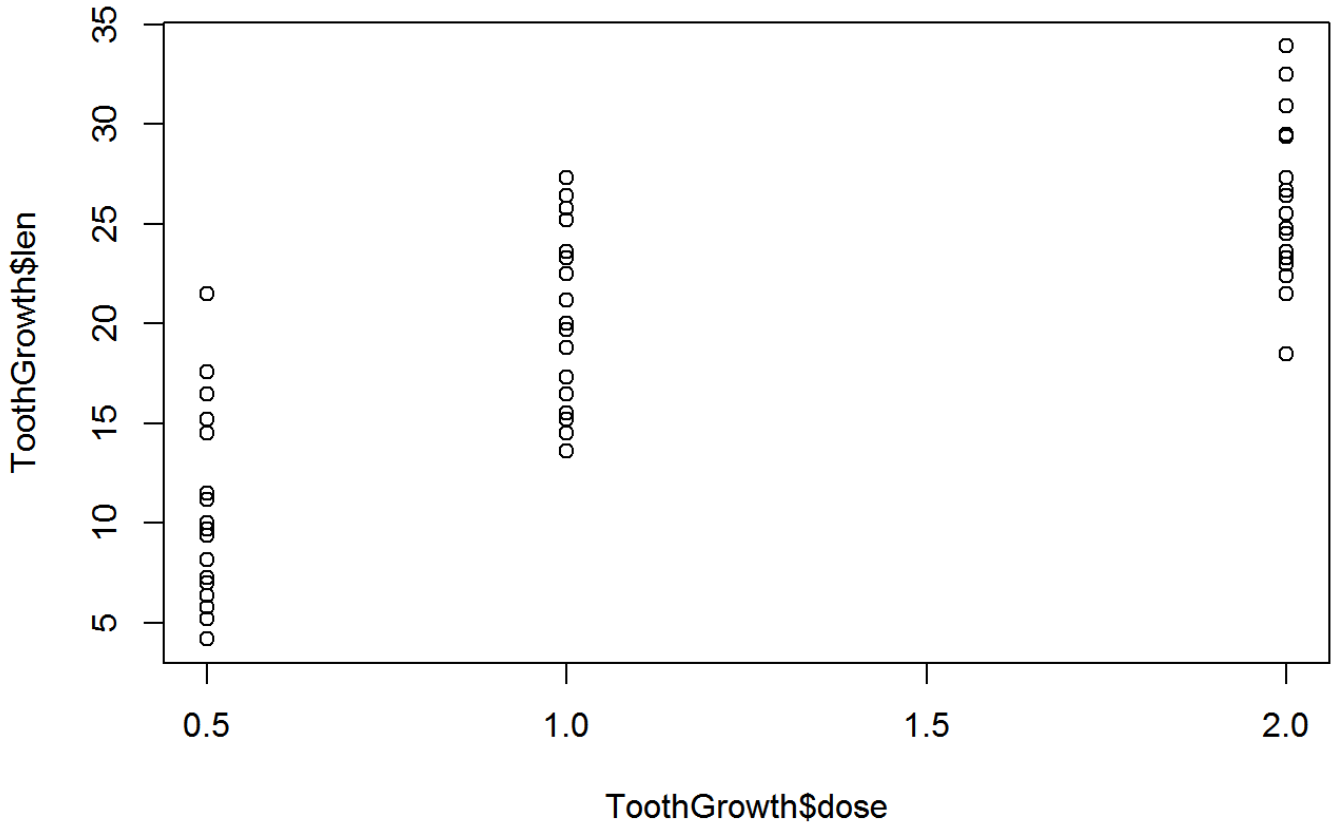


```
boxplot(ToothGrowth$len)
```

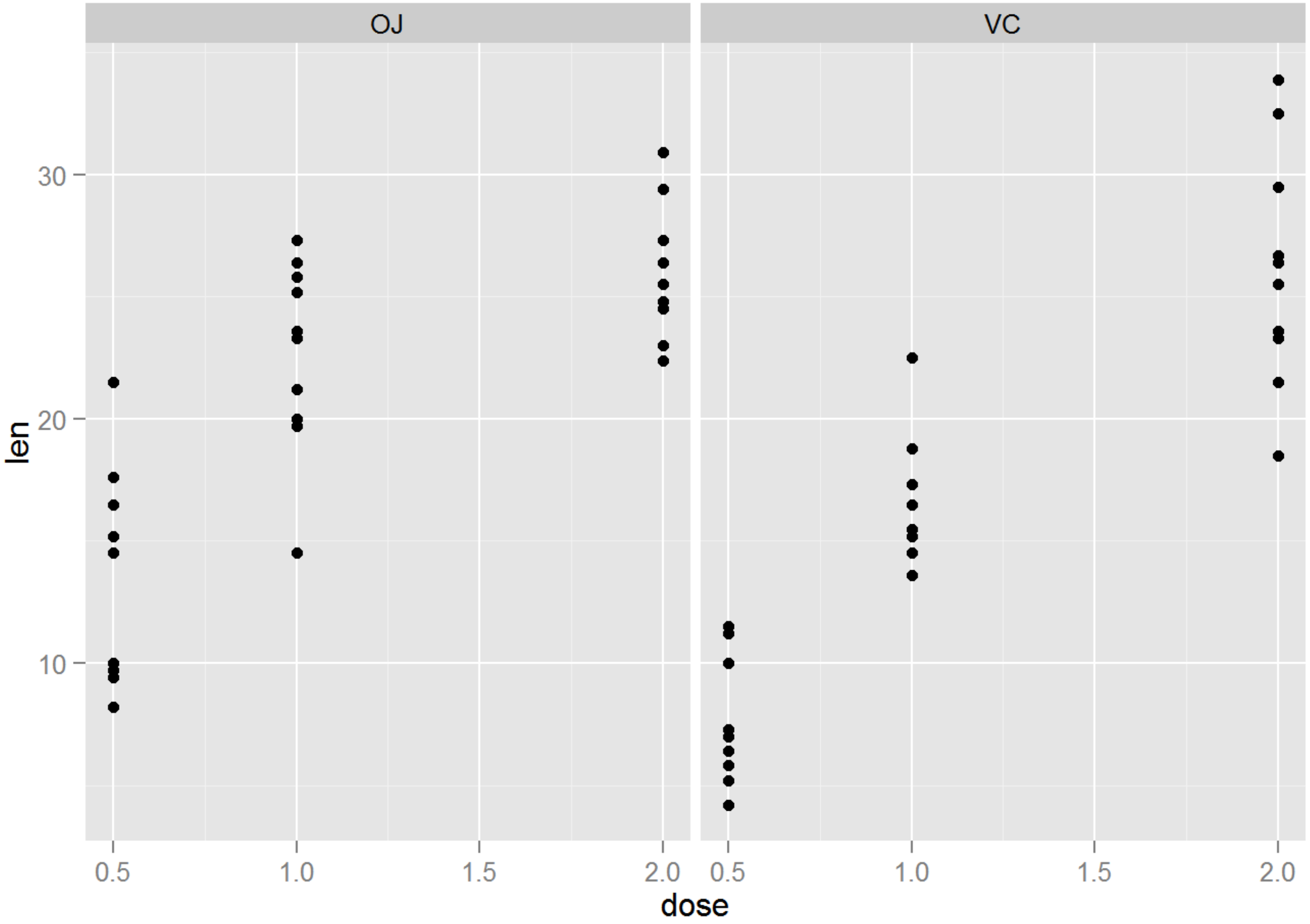


```
plot(ToothGrowth$dose, ToothGrowth$len)

library(ggplot2)
```



```
ggplot(data=ToothGrowth, aes(dose, len)) + geom_bar() + geom_point() + facet_wrap(~supp)
```



Above plots illustrate that length might be increasing with number of doses and supplements, which we validate in below testing

HYPOTHESIS TESTING

We check following null hypotheses

1. There is no significant difference in length with supplements

```
t.test(len ~ supp, paired = F, var.equal = F, 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
```

We can not reject NULL hypthesis with this confidence interval.

2. There is no significant difference in length with dose

```
# COMPARISON WITH 0.5 and 1.0
ToothGrowth1 <- subset(ToothGrowth, ToothGrowth$dose %in% c(0.5,1.0))
ToothGrowth2 <- subset(ToothGrowth, ToothGrowth$dose %in% c(1.0,2.0))
ToothGrowth3 <- subset(ToothGrowth, ToothGrowth$dose %in% c(0.5,2.0))
t.test(len ~ dose, paired = F, var.equal = F, data = ToothGrowth1)
```

```
## 
##  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(len ~ dose, paired = F, var.equal = F, data = ToothGrowth2)
```

```
## 
##  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(len ~ dose, paired = F, var.equal = F, data = ToothGrowth3)
```

```
##
##  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
```

High confidence for alternate hypothesis indicate that we can reject NULL hypothesis.

We made some assumption while checking above hypothesis:-

- There might be other variables affecting data, which is missing in given ToothGrowth dataset.
- There might be mistake in collecting data.
- There might be affect of one dose into others does, if same guinea pig is tries with all doses.
- It might not blind-eyed test i.e. guinea pig might be aware of doses.
- sample population might not be independent/random.