Statistical Inference Project : Part 2

Pradeepta Das

8th November 2020

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from 
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

Part 2: Basic Inferential Data Analysis Instructions

Data Exploration and Visualization

```
data(ToothGrowth)
?ToothGrowth

# Look at the structure of the data
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 ...
```

```
# Look at the first 5 rows of the data
head(ToothGrowth, 5)
```

```
##    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
```

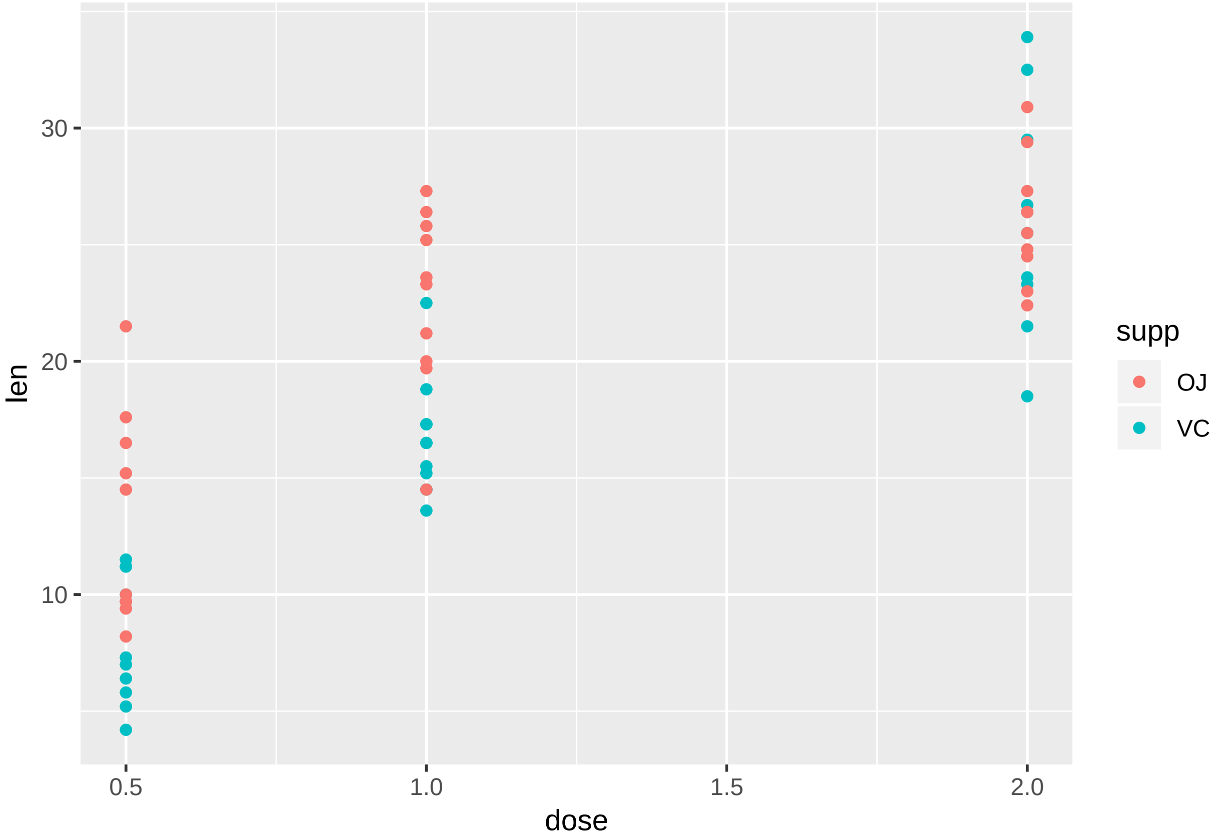
```
# Look at summary the data where supply is Orange Juice
summary(ToothGrowth[ToothGrowth$supp=="OJ",])
```

```
##       len         supp         dose      
##  Min.   : 8.20   OJ:30   Min.   :0.500  
##  1st Qu.:15.53   VC: 0   1st Qu.:0.500  
##  Median :22.70           Median :1.000  
##  Mean   :20.66           Mean   :1.167  
##  3rd Qu.:25.73           3rd Qu.:2.000  
##  Max.   :30.90           Max.   :2.000  
```

```
# Look at summary the data where supply is Ascorbic Acid
summary(ToothGrowth[ToothGrowth$supp=="VC",])
```

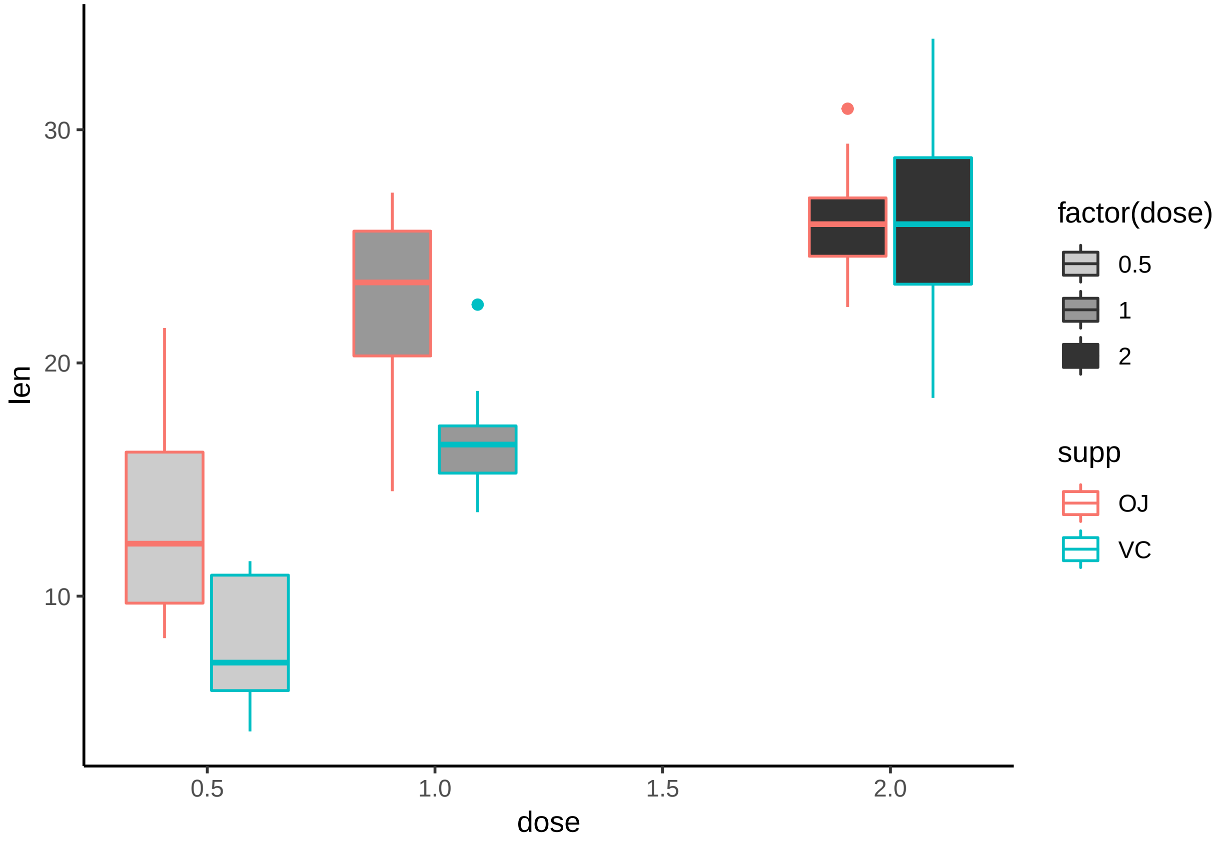
```
##       len         supp         dose      
##  Min.   : 4.20   OJ: 0   Min.   :0.500  
##  1st Qu.:11.20   VC:30   1st Qu.:0.500  
##  Median :16.50           Median :1.000  
##  Mean   :16.96           Mean   :1.167
##  3rd Qu.:23.10           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

```
# Make a plot to look at data graphically
p <- ggplot(data=ToothGrowth, aes(x=dose, y=len))+geom_point(aes(color=supp))
print(p)
```



Looks like for low dose 0.5, there is a clear distinction between VC and OJ.

```
p2 <- ggplot(data=ToothGrowth,
             aes(x=dose, y=len, fill = factor(dose)))+
      geom_boxplot(aes(color=supp)) +
      #scale_fill_brewer(palette="Accent")
      scale_fill_grey(start=0.8, end=0.2) + theme_classic()
print(p2)
```



Basic Summary

From the baove two plots, it is clear that Orange Juice seems to be more effective for the lower doses 0.5 and 1. However, for dose = 2 both seem to do a similar kind of job. In the high dose case also Orange juice has lower variance. So, overall orange juice seems to be doing a better job!

Statistical Analysis

Now let's see if we can justify the same using the stastistical inferences.

Confidence Interval

Comparison by delivery method for the same dosage

Here the null-hypothesis is there is no difference between the delivery methods for 0.5 dose. Similarly other two null hypothesis are for 1 and 2 doses.

```r
t05 <- t.test(ToothGrowth[(ToothGrowth$dose == 0.5) & (ToothGrowth$supp == "OJ"),]$len,
              ToothGrowth[(ToothGrowth$dose == 0.5) & (ToothGrowth$supp == "VC"),]$len,
              paired = FALSE,
              var.equal = FALSE)

t1 <- t.test(ToothGrowth[(ToothGrowth$dose == 1) & (ToothGrowth$supp == "OJ"),]$len,
              ToothGrowth[(ToothGrowth$dose == 1) & (ToothGrowth$supp == "VC"),]$len,
              paired = FALSE,
              var.equal = FALSE)
```

```
t2 <- t.test(ToothGrowth[(ToothGrowth$dose == 2) & (ToothGrowth$supp == "OJ"),]$len,
              ToothGrowth[(ToothGrowth$dose == 2) & (ToothGrowth$supp == "VC"),]$len,
              paired = FALSE,
              var.equal = FALSE)
```

Summary from these 3 tests:

```
summary_dose <- data.frame(
      "p-value" = c(t05$p.value, t1$p.value, t2$p.value),
      "Low.Confidence" = c(t05$conf.int[1], t1$conf.int[1], t2$conf.int[1]),
      "Low.Confidence" = c(t05$conf.int[2], t1$conf.int[2], t2$conf.int[2]),
      row.names = c("Dosage 0.5","Dosage 1","Dosage 2"))
summary_dose
```

```
##                 p.value Low.Confidence Low.Confidence.1
## Dosage 0.5 0.006358607       1.719057         8.780943
## Dosage 1   0.001038376       2.802148         9.057852
## Dosage 2   0.963851589      -3.798070         3.638070
```

p-value adjustment

This doesn't have any meaning here; because we don't have repeated experiments!

```
p.adjust(summary_dose$p.value, method="BH")
```

```
## [1] 0.009537910 0.003115128 0.963851589
```

All these values do not mean anything!

Conclusion

We can reject the null hypothesis for 0.5 dose with 95% confidence. Also the t-confidence interval do not contain zero. The p-value is less than the threshold 0.05 as well. So there definitely is a difference between the delivery methods of Orange Juice ans Ascorbic Acid for the 0.5 dose. And the orange juice delivery method seems to work better. Same is the case for the dose = 1.

However, for the dose = 2 case, with 95% confidence we fail to reject the null hypothesis, i.e. there is no difference in the tooth growth by the delivery methods. Also, we observe p-values more than the treshold of .05 and the confidence levels include 0. So, for dosage of 2 milligrams/day the delivery method doesn't matter.

Assumption: t-distribution for tooth lengths. No other unmeasured factors are affecting tooth length.