R - Statistical Inference Course Project - Part 2

Uday T

27/08/2019

Now in the second portion of the project, 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 2).Provide a basic summary of the data. 3).Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there's other approaches worth considering) 4).State your conclusions and the assumptions needed for your conclusions . ## 1.Load the ToothGrowth data to perform some basic EDA

```
library(datasets)
data("ToothGrowth")
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.1
```

Exploring the contents of the dataset

```
#How the dataset loks like
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
```

```
#Type of data and types of variables
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 ...
```

##2).Provide a basic summary of the data.

```
#Observations and Summary
table(ToothGrowth$dose, ToothGrowth$supp)
```

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

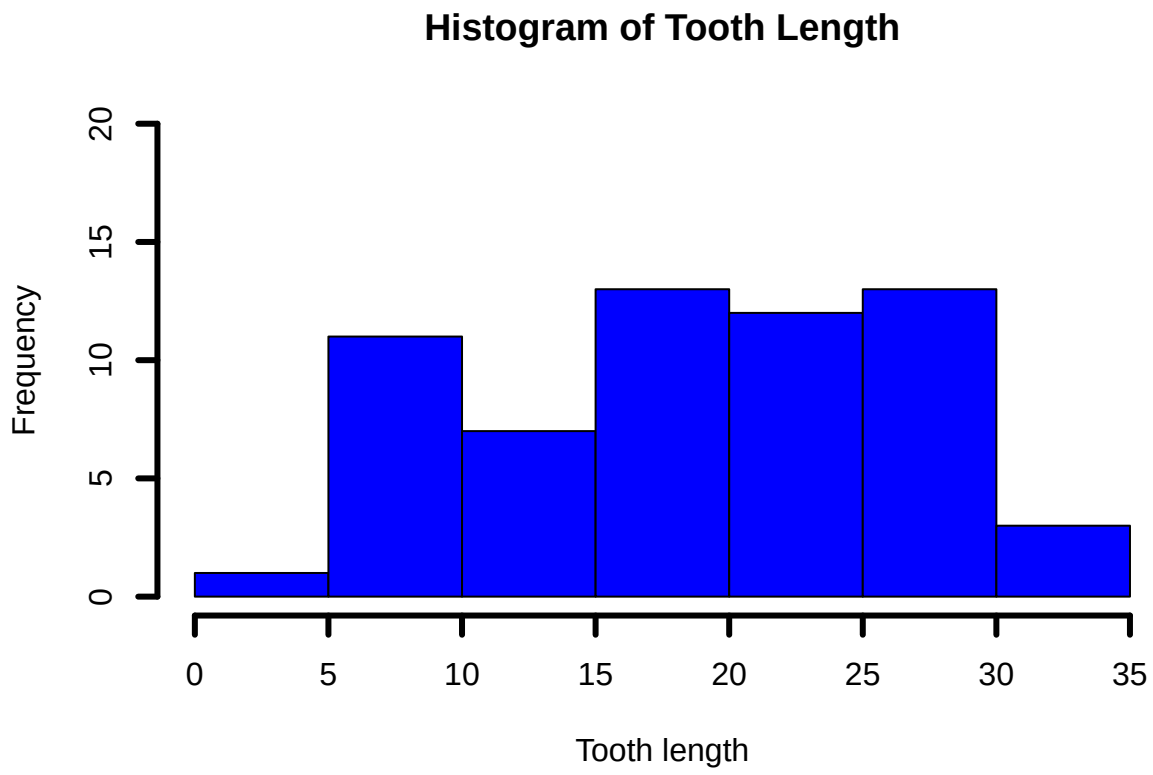
```
print("----------------------------------")
```

```
## [1] "----------------------------------"
```

```
print(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
```

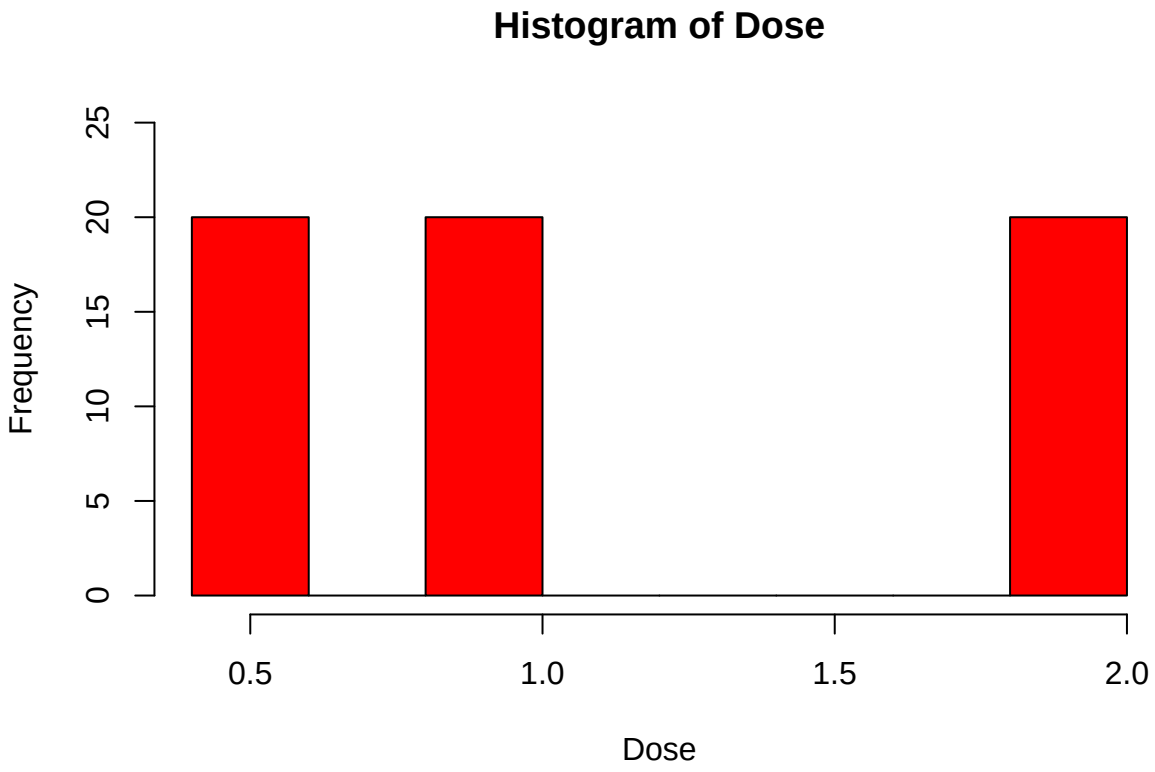
```
#Histogram of Tooth Length
hist(ToothGrowth$len, main = "Histogram of Tooth Length",xlab = "Tooth length", ylab = "Frequency",
     ylim = c(0,20),col  = "blue", lwd  = 3)
```

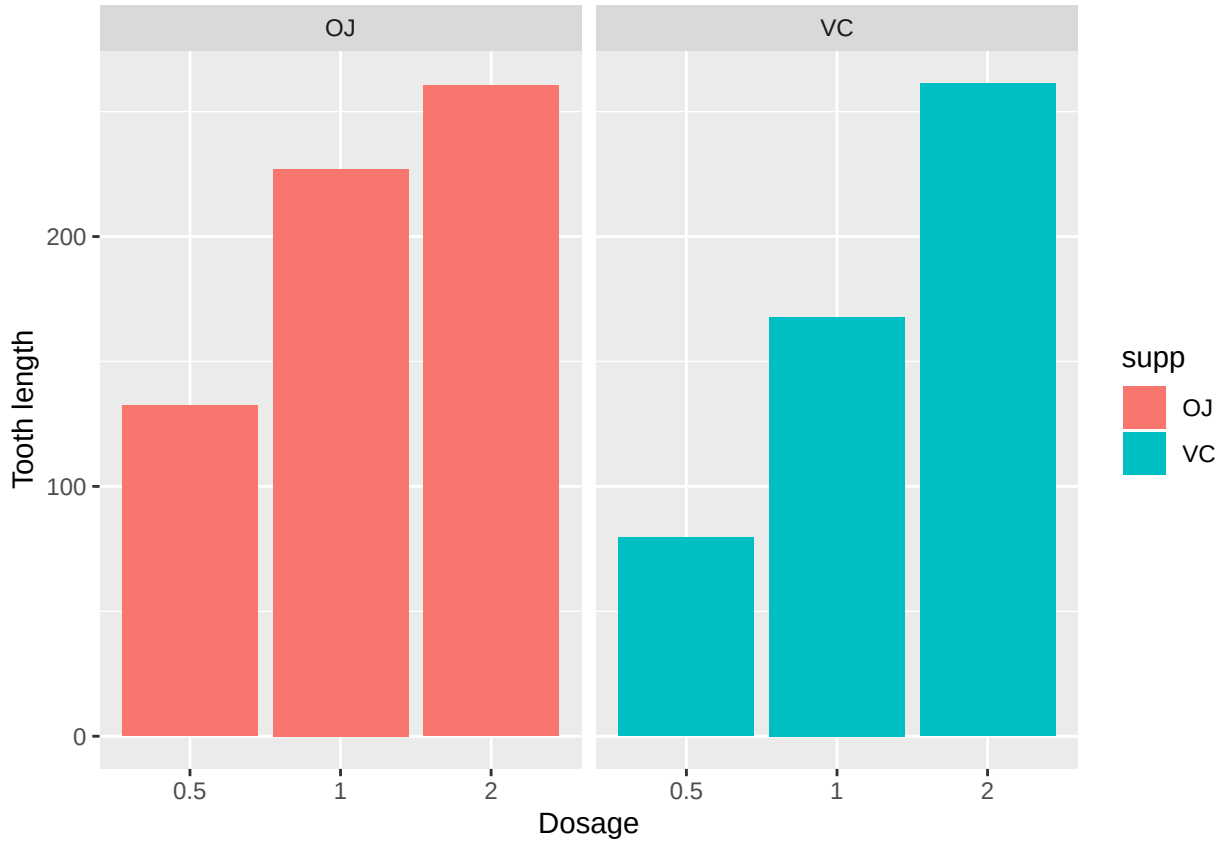



```
#Histogram of Dose
hist(ToothGrowth$dose, main = "Histogram of Dose", xlab = "Dose", ylab = "Frequency",
     ylim = c(0,25), col  = "red",slwd  = 1)
```

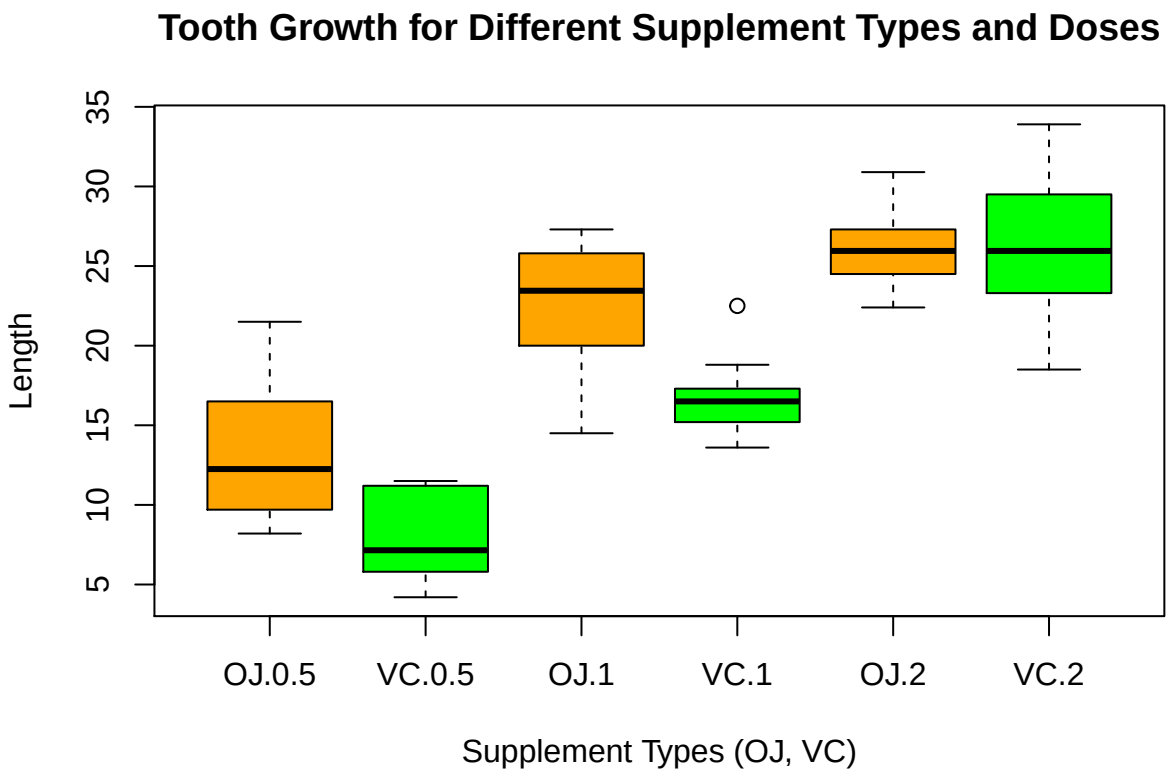
```
## Warning in plot.window(xlim, ylim, "", ...): "slwd" is not a graphical
## parameter
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## "slwd" is not a graphical parameter

## Warning in axis(1, ...): "slwd" is not a graphical parameter

## Warning in axis(2, ...): "slwd" is not a graphical parameter
```






```
boxplot(len~supp+dose, data = ToothGrowth,
        main = "Tooth Growth for Different Supplement Types and Doses",
        xlab = "Supplement Types (OJ, VC)", ylab = "Length",
        col  = c("orange","green"))
```



– The Tooth length varies from 0 to 35, but most of the values range from 5 to 30 – There are 3 different dose types, 0.2, 1.0 and 2.0 wit equal frequencies, i.e 20 each! – There are two different dosages, OJ and VC - Orange Juice and Vitamin C with 0.5 dose having least frequency and 2.0 having highest frequecy for both dosages

Lets now conduct Hypothesis tests on Toothlength and Dose data,

3. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose.

```
# Let's perform t-test
t.test(len ~ supp, paired = FALSE, var.equal = FALSE, 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
```

Let's compare Tooth Growth by dose

```
t.test(len ~ supp, ToothGrowth[ToothGrowth$dose == .5, ])
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.719057 8.780943
## sample estimates:
## mean in group OJ mean in group VC
##            13.23             7.98
```

```
t.test(len ~ supp, ToothGrowth[ToothGrowth$dose == 1, ])
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.802148 9.057852
## sample estimates:
## mean in group OJ mean in group VC
##            22.70            16.77
```

```
t.test(len ~ supp, ToothGrowth[ToothGrowth$dose == 2, ])
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = -0.046136, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean in group OJ mean in group VC
##            26.06            26.14
```

So now we move on to dosage wise analysis, We'll split the data by range of values in Dosages:

```
dose.one <- subset(ToothGrowth, dose %in% c(0.5, 1.0))
dose.two <- subset(ToothGrowth, dose %in% c(1.0, 2.0))
dose.three <- subset(ToothGrowth, dose %in% c(0.5, 2.0))
```

Our null hypothesis is that there is no difference in tooth length between the different dosages.

```
t.test(len ~ dose, paired=F, var.equal=F, data=dose.one)
```

```
##
##  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=dose.two)
```

```
##
##  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=dose.three)
```

```
##
##  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
```

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

##Conclusions:

The Tooth length varies from 0 to 35, but most of the values range from 5 to 30 There are 3 different dose types, 0.2, 1.0 and 2.0 wit equal frequencies, i.e 20 each! There are two different dosages, OJ and VC - Orange Juice and Vitamin C with 0.5 dose having least frequency and 2.0 having highest frequecy for both dosages

The p-value is greater than 0.05 and the confidence interval of the test contains zero so we can say that supplement types seems to have no impact on Tooth growth based on this test.

We also find a 95% confidence interval of -0.171, 7.571 for mean(OJ)-mean(VC). Our p value is 0.06063, which is not significant. Therefore, we would not reject the null hypothesis.

Only the dose with value 1 is showing a significant impact on the Tooth Growth, This might also tell that the tooth growth has very less correlation or no correlation with dosage and dosage type, With a p-value = 0.06 and having zero in the confidence interval we can not reject the null hypothesis.

DOSE ONE - We find a 95% confidence interval of -11.98, -6.27 for mean(0.5)-mean(1.0) with a p value of 1.268e-07. Therefore, we reject the null hypothesis.

DOSE TWO - We find a 95% confidence interval of -8.99, -3.73 for mean(1.0)-mean(2.0) with a p value of 1.906e-05. Therefore, we reject the null hypothesis.

DOSE THREE - We find a 95% confidence interval of -18.15, -12.83 for mean(0.5)-mean(2.0) with a p value of 4.398e-14. Therefore, we reject the null hypothesis.

The length increases with increase in dosage for both the supplements.In all three tests, we see that none of the 95% CL intervals include zero and they are always on the negative side. Therefore, we can claim, with 95% CL, that the is a positive correlation between the two variables.

We can conclude that supplement delivery method has no effect on Tooth Length even though they are positively correlated.
