Peer Graded Stat. Inf. Part 2

Dasarath S

23/10/2020

Brief

In this part 2 of the project, we are going to use ToothGrowth data from the package - R datasets to do simple inferential analysis.

```
knitr::opts_chunk$set(echo = TRUE)
```

We need `ggplot2` to successfully perform this analysis.

```
library(ggplot2)
```

Part 2: Doing Simple Inferential Data Analysis on ToothGrowth dataset

1. Loading the ToothGrowth data using data() function

```
data(ToothGrowth)
```

2. Give an outlook of the data using summary() and do some simple exploratory analysis on the same

```
TG<-ToothGrowth
str(TG)
```

```
## '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 ...
```

```
length<-ToothGrowth$len
dosage<-as.character(TG$dose)
supplement <- as.character(TG$supp)
summary(length)
```

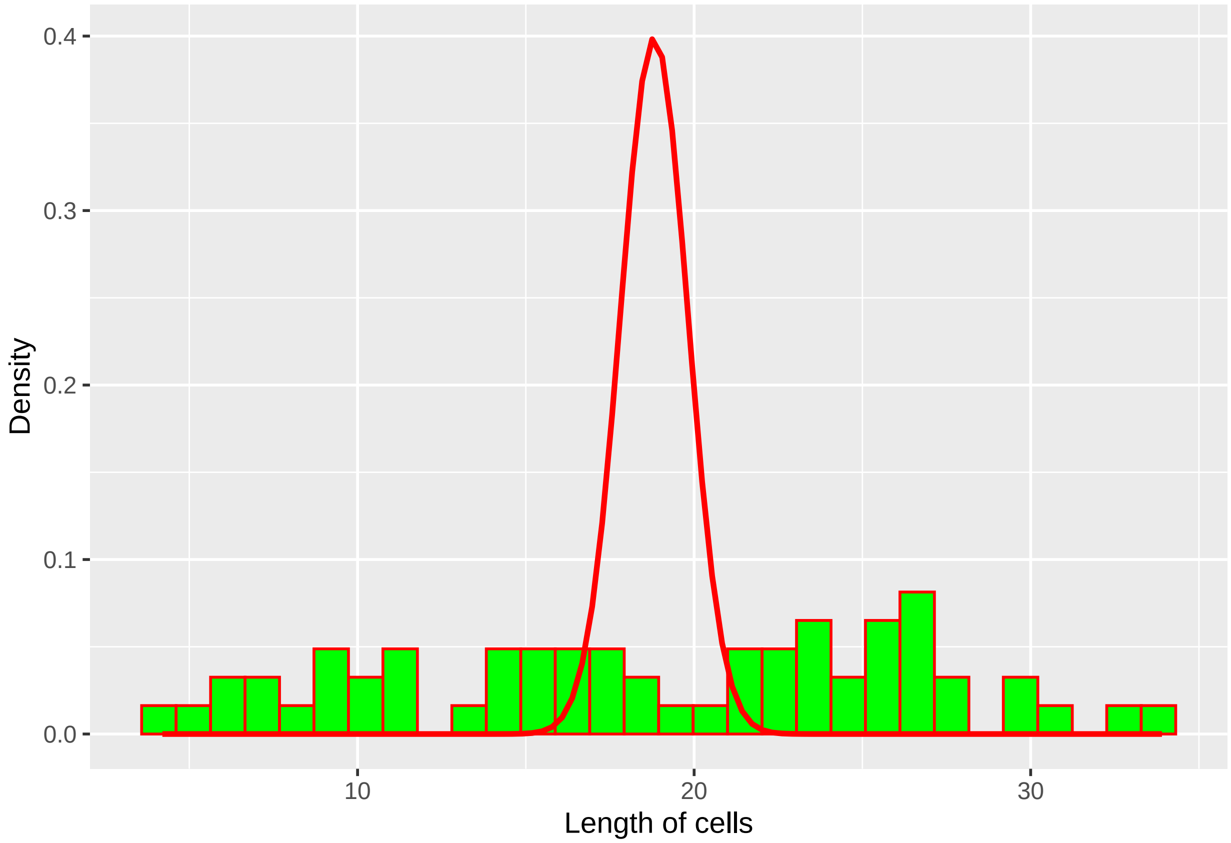
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.20   13.07   19.25   18.81   25.27   33.90
```

Looking at the documentation of this dataset on R, we come to know that this data is the length of cells in tooth called odontoblasts which are the reason for the growth of teeth. The data comes from 60 guinea pigs out of which 10 of them were given 3 different doses. Two different ways of delivering Vitamin C - one via Orange Juice(OJ) and the other via ascorbic acid(VC). Also 3 different dosages of vitamin C - half, one and two mg per day were administered to each group.

```
# Plotting hist of length so as to see how its distributed
Plot <- ggplot(TG,aes(length))
Plot + geom_histogram(aes(y = ..density..),colour="red",fill="green")+
       stat_function(fun=dnorm, color = "red", sd=sqrt(var(TG$len)),
       args=list( mean=mean(length)),
       size = 1.0, geom="line")+ ylab("Density") + scale_x_continuous("Length of cells")
```

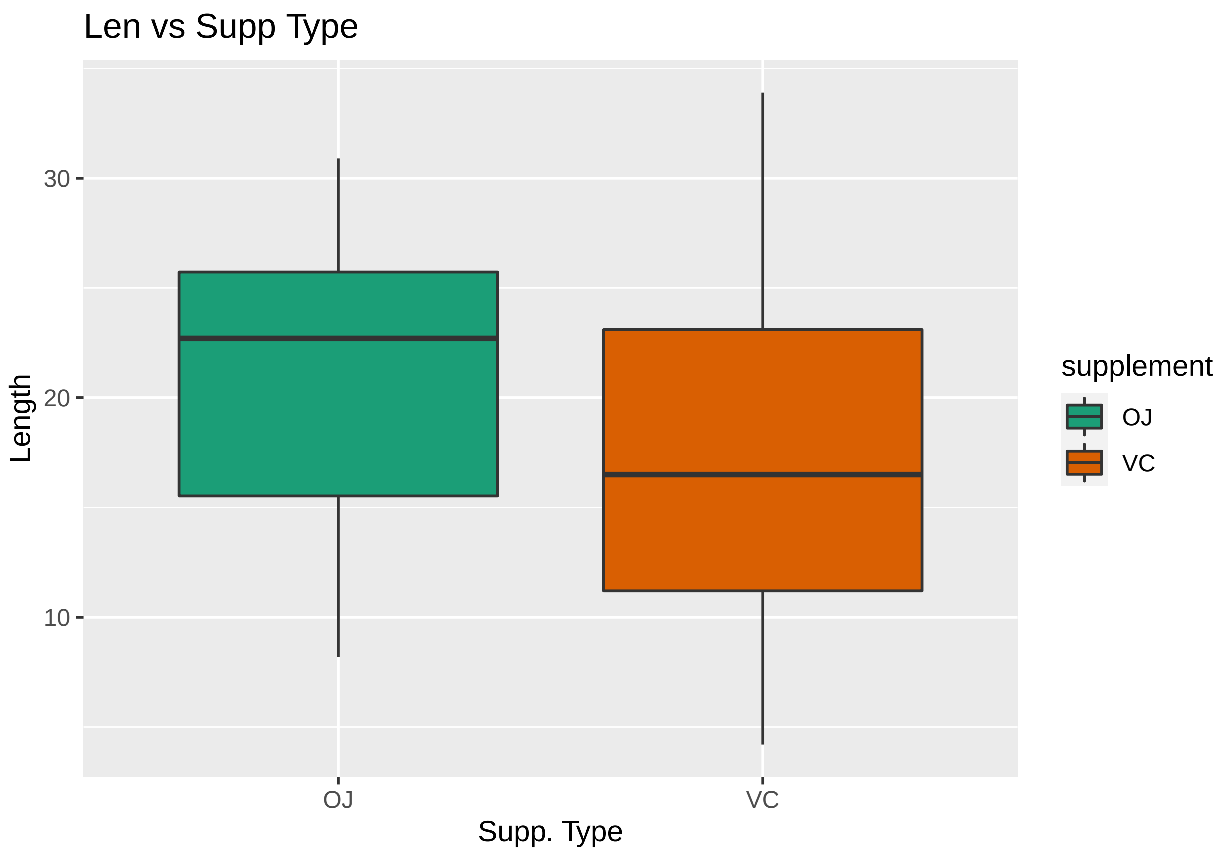
```
## Warning: Ignoring unknown parameters: sd
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



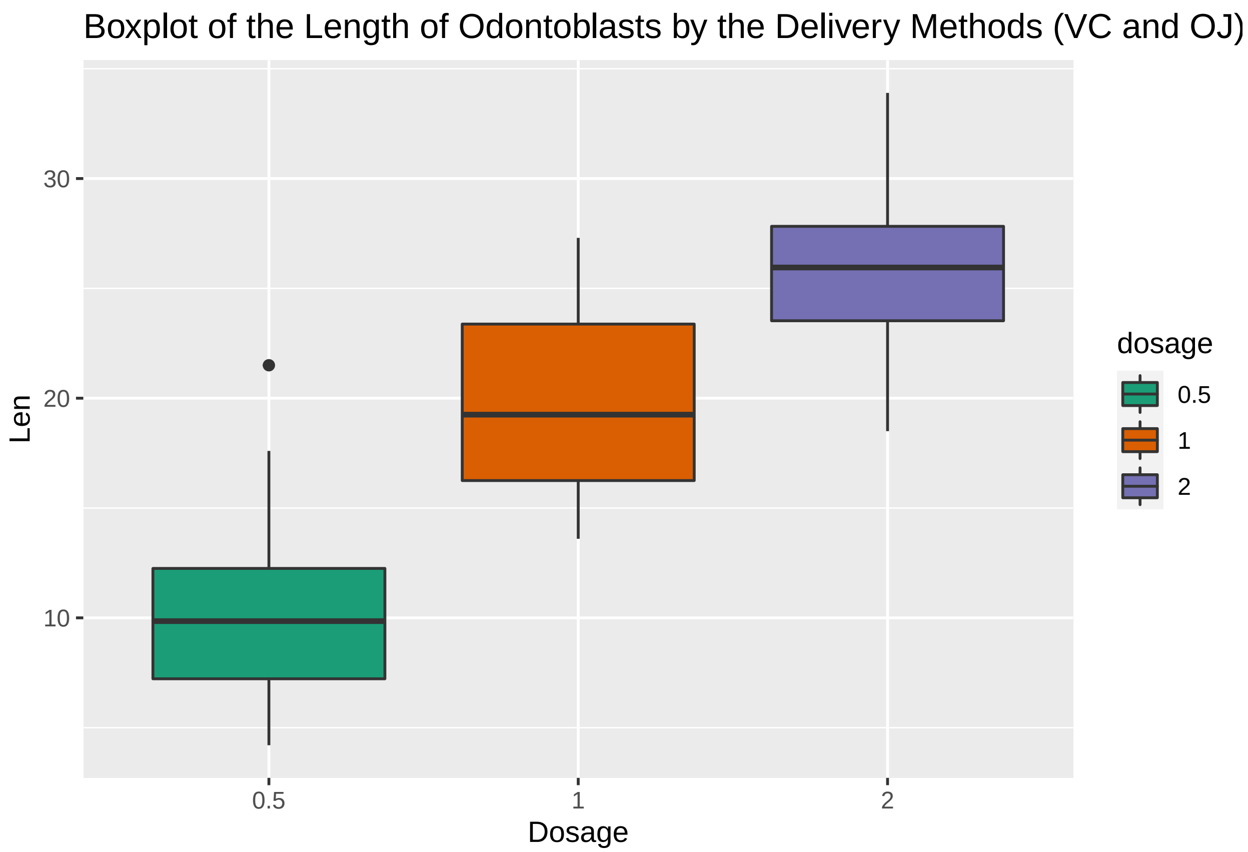
We can see that the green histogram bars' height is small compared to the computed normal distribution which is marked as red. The reason for this can be because of different groups of pigs getting different levels of the vitamin C. We have to do further exploration of length vs the remaining variables.

```
# Seeing how length and supplement are related with the help of a boxplot
Plot <- ggplot(TG,aes(supplement,length))
Plot+geom_boxplot(aes(fill=supplement))+
      ggtitle("Len vs Supp Type")+
      xlab("Supp. Type")+
      ylab("Length") + scale_fill_brewer(palette="Dark2")
```



From the boxplots its clearly evident that the median of OJ isa bit higher than the median of VC when plotted with length. Also another observation is that since the top half of OJ is smaller compared to the bottom half, it means its a bit skewed towards left in the distribution, whereas VC is more centrally distributed.

```
Plot <- ggplot(TG,aes(dosage,length))
Plot + geom_boxplot(aes(fill=dosage))+
      ggtitle("Boxplot of the Length of Odontoblasts by the Delivery Methods (VC and OJ)")+
      xlab("Dosage")+
      ylab("Len") + scale_fill_brewer(palette="Dark2")
```



An interesting observation from the above boxplots where length vs doses are plotted is that as dosage median increases we also can see that the length increases. distributions are approximately normal as the medians are centred, but dose have similar variability.

3. Contrast tooth growth by dosage and supplement by using the methods of confidence intervals and/or hypothesis tests.

Setting hypothesis for the initial test:

```
# Use the function called t-test() on each dosage group to print the corresponding p-Values
dosage_level<-levels(factor(TG$dose))
rejection<-paste(" ")
no_rejection<-paste(" ")
for (each_level in dosage_level){
    resulting_value<-t.test(len ~ supp, TG[TG$dose == each_level, ])
    print(paste("For the dosage group",as.character(each_level),
                "the resulting t.test value is: "))
    print(resulting_value)
    ifelse(resulting_value$p.value<0.05,
           print(paste("We are rejecting the null hypothesis for the dosage group"
                       ,as.character(each_level)," as the p-value < 5%")),
           print(paste("We arent rejecting the null hypothesis for the dosage group"
                       ,as.character(each_level)," as the p-value > 5%")))
    "\n"
    ifelse(resulting_value$p.value<0.05,
           rejection<-paste(as.character(each_level), " & ", rejection),
           no_rejection <-paste(as.character(each_level), " & ", no_rejection))
    }
```

```
## [1] "For the dosage group 0.5 the resulting t.test value is: "
##
##  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
##
## [1] "We are rejecting the null hypothesis for the dosage group 0.5  as the p-value < 5%"
## [1] "For the dosage group 1 the resulting t.test value is: "
##
##  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
##
## [1] "We are rejecting the null hypothesis for the dosage group 1  as the p-value < 5%"
## [1] "For the dosage group 2 the resulting t.test value is: "
##
##  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
##
## [1] "We arent rejecting the null hypothesis for the dosage group 2  as the p-value > 5%"
```

4. Point out the assumptions you based your conclusions on and the conclusions itself

Below are the conclusions from the analysis:

- From our first set of exploratory analysis, it seems like the most likely factor for the length of the tooth growth cells is the dosage of vitamin C rather than the type of supplement.
- We can say that true average difference of length of teeth growth cells due to orange juice and ascorbic acid against 3 different dosages with a 95% confidence because the t-test results have an interval of 95% confidence.

- We aren't rejecting the difference with dosage levels with 2 & given, whereas we will be rejecting the null hypothesis H_0 with dosage 1 & 0.5 & given for the difference in supplement, ascorbic acid and orange juice.
- The assumptions we make here for the above conclusions is the normal and independent distribution of data for the t-test condition.