MDSC-206-Assignment-IV-20230

jyotin padhi

26 February 2021

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3 v purrr 0.3.4  
## v tibble 3.0.6 v dplyr 1.0.3  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(lattice)  
library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

## The following object is masked from 'package:purrr':  
##   
## some

library(gridExtra)

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

library(funModeling)

## Loading required package: Hmisc

## Loading required package: survival

## Loading required package: Formula

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':  
##   
## src, summarize

## The following objects are masked from 'package:base':  
##   
## format.pval, units

## funModeling v.1.9.4 :)  
## Examples and tutorials at livebook.datascienceheroes.com  
## / Now in Spanish: librovivodecienciadedatos.ai

#loading dataset  
data<-ToothGrowth  
#looking to the first 6 rows  
head(data)

## 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

#getting the column names of the dataset  
names(data)

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

#seeing the dimension of dataset  
dim(data)

## [1] 60 3

#getting the summary of the dataset  
summary(data)

## 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

#check for na values  
sum(is.na(data))

## [1] 0

#check for duplicated values  
sum(duplicated(data))

## [1] 5

#converting the numerical tyoe to factorial in order to apply anova test  
data$dose <- as.factor(data$dose)  
glimpse(data)

## Rows: 60  
## Columns: 3  
## $ len <dbl> 4.2, 11.5, 7.3, 5.8, 6.4, 10.0, 11.2, 11.2, 5.2, 7.0, 16.5, 16...  
## $ supp <fct> VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC...  
## $ dose <fct> 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 1, 1, 1, 1, ...

attach(data)  
by(data,INDICES=supp,FUN=summary)

## supp: OJ  
## len supp dose   
## Min. : 8.20 OJ:30 0.5:10   
## 1st Qu.:15.53 VC: 0 1 :10   
## Median :22.70 2 :10   
## Mean :20.66   
## 3rd Qu.:25.73   
## Max. :30.90   
## ------------------------------------------------------------   
## supp: VC  
## len supp dose   
## Min. : 4.20 OJ: 0 0.5:10   
## 1st Qu.:11.20 VC:30 1 :10   
## Median :16.50 2 :10   
## Mean :16.96   
## 3rd Qu.:23.10   
## Max. :33.90

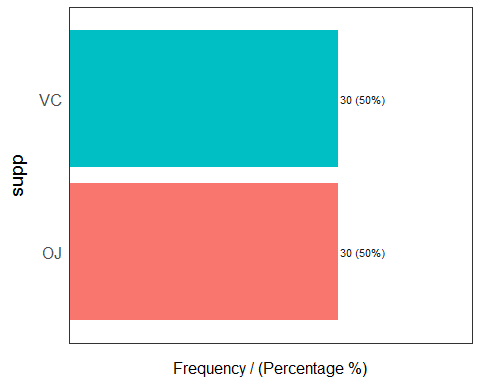
by(data,INDICES=dose,FUN=summary)

## dose: 0.5  
## len supp dose   
## Min. : 4.200 OJ:10 0.5:20   
## 1st Qu.: 7.225 VC:10 1 : 0   
## Median : 9.850 2 : 0   
## Mean :10.605   
## 3rd Qu.:12.250   
## Max. :21.500   
## ------------------------------------------------------------   
## dose: 1  
## len supp dose   
## Min. :13.60 OJ:10 0.5: 0   
## 1st Qu.:16.25 VC:10 1 :20   
## Median :19.25 2 : 0   
## Mean :19.73   
## 3rd Qu.:23.38   
## Max. :27.30   
## ------------------------------------------------------------   
## dose: 2  
## len supp dose   
## Min. :18.50 OJ:10 0.5: 0   
## 1st Qu.:23.52 VC:10 1 : 0   
## Median :25.95 2 :20   
## Mean :26.10   
## 3rd Qu.:27.82   
## Max. :33.90

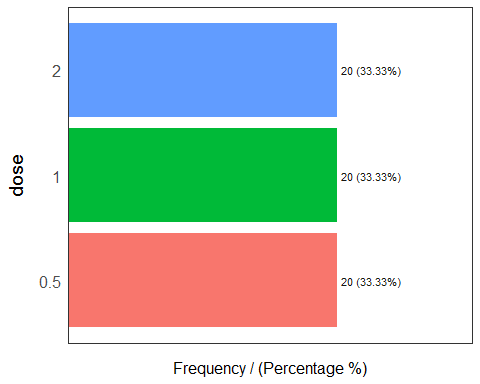
#here we can see the individual summaries of the catogorical variable

## EDA

freq(data)



## supp frequency percentage cumulative\_perc  
## 1 OJ 30 50 50  
## 2 VC 30 50 100



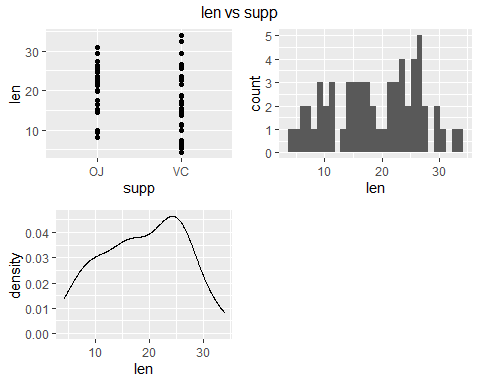
## dose frequency percentage cumulative\_perc  
## 1 0.5 20 33.33 33.33  
## 2 1 20 33.33 66.66  
## 3 2 20 33.33 100.00

## [1] "Variables processed: supp, dose"

#here we can see that there are equal number classes in both sup and dose

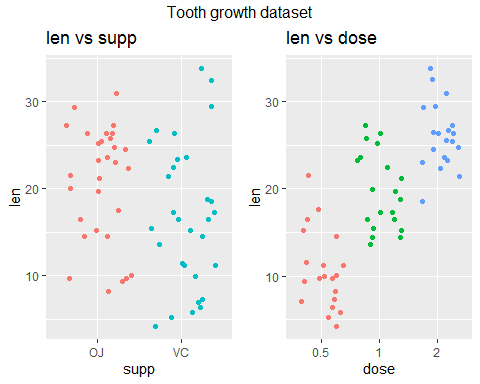
grid.arrange(ggplot(data, aes(x=supp, y=len))+geom\_point(),  
 ggplot(data, aes(len))+geom\_histogram(),  
 ggplot(data, aes(len))+geom\_density(),  
 nrow=2, top='len vs supp')

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

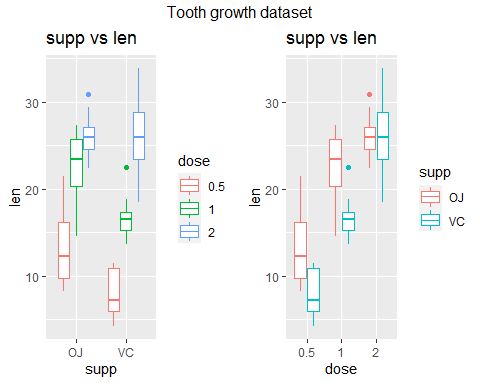


#some visualizations to know the behaviour of length

grid.arrange(ggplot(data)+  
 aes(supp,len,color=supp)+  
 geom\_jitter()+  
 theme(legend.position="none")+  
 labs(title="len vs supp"),ggplot(data)+  
 aes(dose,len,color=dose)+  
 geom\_jitter()+  
 theme(legend.position="none")+  
 labs(title="len vs dose"), nrow=1, top='Tooth growth dataset' )

 It's clear here that the toth lenghts are lower for the ones with lower dose and increases as dose increases

grid.arrange(ggplot(data)+  
 aes(supp,len,color=dose)+  
 geom\_boxplot(na.rm = TRUE)+ # geom\_point(position = "jitter")  
 theme(legend.position="right")+  
 labs(title="supp vs len"),  
  
ggplot(data)+  
 aes(dose,len,color=supp)+  
 geom\_boxplot(na.rm = TRUE)+ # geom\_point(position = "jitter")  
 theme(legend.position="right")+  
 labs(title="supp vs len"),  
  
 nrow=1, top='Tooth growth dataset')

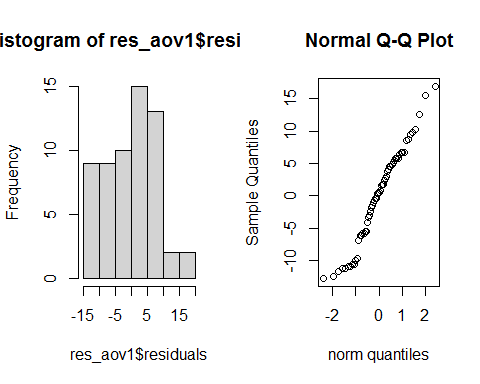
 here we can see that the boxplots tells that the lenghts is more if the dosage is more and if given through VC.

Assumptions for anova are: 1) **Normality** of residuals 2) **Equality** of variances

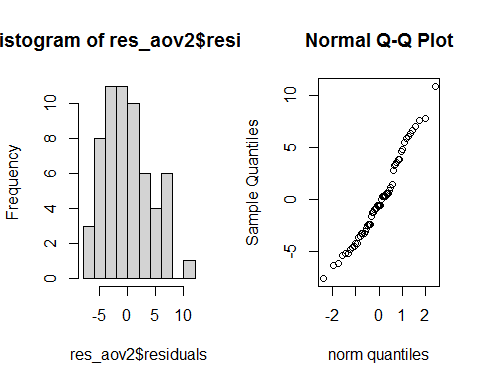
**Normality**

#using aov function to perform the anova test  
res\_aov1 <- aov(len ~ supp,data=data)  
res\_aov2 <- aov(len ~ dose,data=data)

#seeing the normality of residuals visually  
par(mfrow=c(1,2)) # combine plots  
  
# histogram  
hist(res\_aov1$residuals)  
qqnorm(res\_aov1$residuals,xlab = "norm quantiles")

 Since we can see that the model with supp seems to assume normality and can b further confirmed with shapiro test

par(mfrow=c(1,2)) # combine plots  
  
# histogram  
hist(res\_aov2$residuals)  
qqnorm(res\_aov2$residuals,xlab = "norm quantiles")

 Since we can see that the model with dose seems to assume normality and can b further confirmed with shapiro test

shapiro.test(res\_aov1$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: res\_aov1$residuals  
## W = 0.96949, p-value = 0.1378

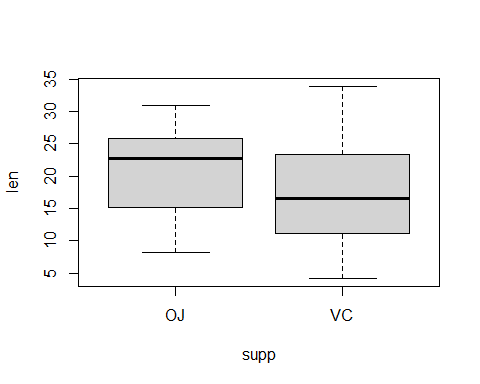
shapiro.test(res\_aov2$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: res\_aov2$residuals  
## W = 0.96731, p-value = 0.1076

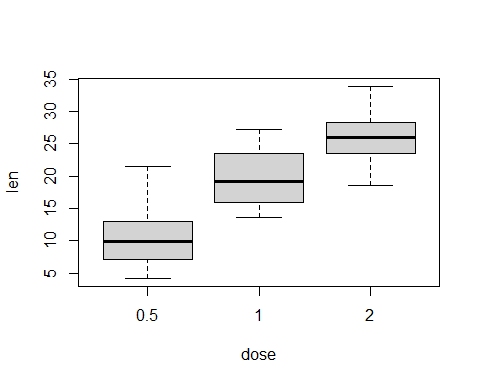
Since the p-value is greater than significance level that is 0.05 accept the null hyothesis that it assumes normality for both supp and dose

**Equality of variances**

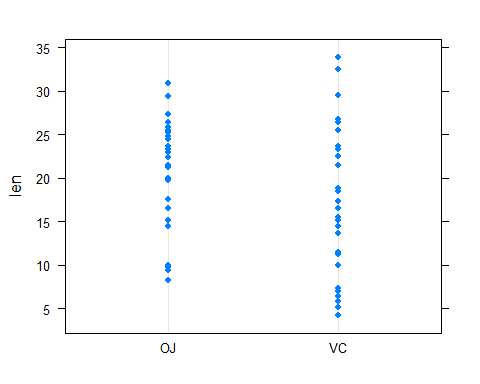
boxplot(len ~ supp, data=data)



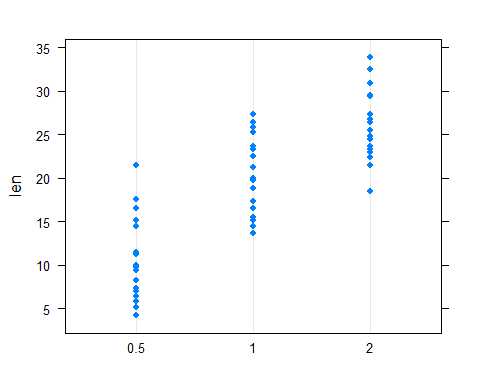
boxplot(len ~ dose, data=data)

 Here we can see visally that the variance of the dose and supp are almost equal and can be further proved statistically by levence test

dotplot(len ~ supp, data=data)



dotplot(len ~ dose, data=data)

 Here we can see visally that the variance of the dose and supp are almost equal and can be further proved statistically by levence test

Testing Hypothesis for Levense Test:

Variances are equal at least one varaince is different

# Levene's test from car library  
leveneTest(len ~ supp, data=data)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 1 1.2136 0.2752  
## 58

leveneTest(len ~ dose, data=data)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 2 0.6457 0.5281  
## 57

since the p-values are greater than significance level i.e 0.05 we accept the null hypothesis that the variances are equal

group\_by(data, supp) %>%   
 summarise(  
 mean = mean(len,na.rm=TRUE),  
 sd = sd(len,na.rm=TRUE)  
 )

## # A tibble: 2 x 3  
## supp mean sd  
## \* <fct> <dbl> <dbl>  
## 1 OJ 20.7 6.61  
## 2 VC 17.0 8.27

group\_by(data, dose) %>%   
 summarise(  
 mean = mean(len,na.rm=TRUE),  
 sd = sd(len,na.rm=TRUE)  
 )

## # A tibble: 3 x 3  
## dose mean sd  
## \* <fct> <dbl> <dbl>  
## 1 0.5 10.6 4.50  
## 2 1 19.7 4.42  
## 3 2 26.1 3.77

Above two we can see the summaries of the dose and supp variables

# One-way ANOVA

## Testing Hypothesis

The null and alternative hypothesis of an ANOVA are: - The three species are equal in terms of tooth lenghts - At least one mean is different.

res\_aov1 <- aov(len ~ supp, data = data)  
summary(res\_aov1)

## Df Sum Sq Mean Sq F value Pr(>F)   
## supp 1 205 205.35 3.668 0.0604 .  
## Residuals 58 3247 55.98   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Since the p-value is greater than the significance level 0.05 we acceot the null hypothesis that the means of OJ and VC are same

res\_aov2 <- aov(len ~ dose, data = data)  
summary(res\_aov2)

## Df Sum Sq Mean Sq F value Pr(>F)   
## dose 2 2426 1213 67.42 9.53e-16 \*\*\*  
## Residuals 57 1026 18   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Since the p-value is lesser than the significance level we reject the null hypothesis (i.e) the means of the factors of doses are not equal

# Two-Way ANOVA

Since each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice or ascorbic acid (a form of vitamin C and coded as VC). We can apply two-way anova test to test the following

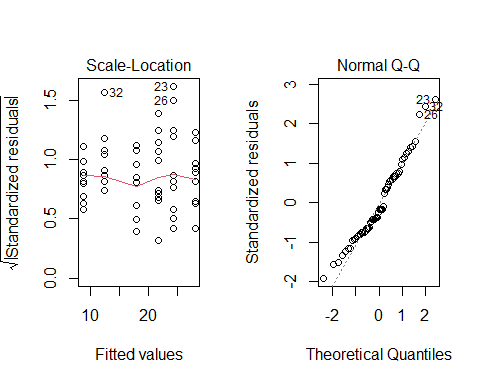
## Testing Hypothesis

The null and alternative hypothesis of an ANOVA are:

* The three species are equal in terms of tooth length
* At least one mean is different.

anova\_two\_way <- aov(len ~ supp+dose,data=data)

par(mfrow=c(1,2)) # combine plots  
plot(anova\_two\_way,3)  
plot(anova\_two\_way,2)



shapiro.test(anova\_two\_way$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: anova\_two\_way$residuals  
## W = 0.96168, p-value = 0.05687

Since the shapiro test's value is greater tahn significance level we can conlcude that the reisfuals are normal

leveneTest(len ~ supp\*dose, data=data)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 5 1.7086 0.1484  
## 54

Since the p-value is graeter than the significance evel 0.001 we can conclude that the variances are equal

summary(anova\_two\_way)

## Df Sum Sq Mean Sq F value Pr(>F)   
## supp 1 205.4 205.4 14.02 0.000429 \*\*\*  
## dose 2 2426.4 1213.2 82.81 < 2e-16 \*\*\*  
## Residuals 56 820.4 14.7   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Since the p-value is lesser than the significance level we can reject null hypothesis and conclude that there is a significane difference between the population mean across the factors