Factorial ANOVA Documentation

Here we are going to conduct Factorial ANOVA test using a built in Data set in R. This data set is called ToothGrowth data set. It contains the length of tooths of 60 Guinea Pigs. The data set has three columns. Namely Len (Length of tooth), supp (supplement provided- two categories Orange Juice (OJ) and Ascorbic Acid (VC)) and dose (dosage of medication-having three categories- 0.5 mm, 1.0 mm, and 2.0 mm)

Note:

Here we consider the Len or tooth length column as response variable. This is continuous which satisfies our assumption 1. We also consider the supp and dose columns as predictors. To conclude the test, the Predictors must be categorical having levels. Here the supp has 2 levels namely OJ and VC. On the other hand, dose column has 3 levels. So, this design is an example of 2*3 factorial design.

Goal: Here we have two goals in this project. They are-

- Is there any change of average length of tooth across different dosage levels?
- Is there a Statistically significant value of mean for ways supplement are provided?

We start this project by installing and loading required packages for this project and loading the in-built ToothGrowth data set in R environment.

```
# Installing and Loading required packages
install.packages("tidyverse")
install.packages("car")
install.packages("emmeans")
library(tidyverse)
library(car)
library(emmeans)
```

In the below code we use the data function to load the in-built data in R environment.

```r
# Loading the in-built data set Tooth Growth in R environment
data("ToothGrowth")
toothData = ToothGrowth
head(toothData,5)
```

	len <dbl>	supp <fctr>	dose <dbl>
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

5 rows

We export the data set so that it can be used as an external file

```r
# Exporting the data as csv file in the same directory where we work for future use
write_csv(toothData, file='Tooth_Growth_Data.csv')
```

Now we perform basic data understanding steps. This includes-

- Getting the column names
- Getting number of rows and column
- Getting the structure
- Getting the summary

```r
# Getting the column names of the tooth-Data data set
colnames(toothData)
```

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

```r
# Getting number of rows and column of tooth-Data data set
nrow(toothData)
```

```
[1] 60
```

```
ncol(toothData)
```

```
[1] 3
```

```
# Getting the structure of the data frame
str(toothData)
```

```
'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 ...
```

```
# Getting Statistical summary for the data frame
summary(toothData)
```

```
      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
```

Now here we are interested in conducting Factorial ANOVA test. For this we must have two conditions to satisfy. They are-

- The **response variable is Continuous.**
- The **independent variables are categorical.**

But in our case the **dose column dose is numerical.** We have to convert it into categorical variables having 3 levels. We say **0.5 mm dose as low, 1.0 mm dose as medium** and **2.0 mm dose as high** dose. The below code do the same for us.

```
# Converting dose column into categorical variable
toothData$dose = factor(toothData$dose,levels=c(0.5,1.0,2.0),labels=c("low","medium",
"high"))
str(toothData)
```

```
'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: Factor w/ 3 levels "low","medium",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Now before moving to the main works let us make the contingency table for supp and dose from this toothData data set. This will help to

- By observing the respective frequencies in all combinations.
- It also helps to observe if there is any pattern in the data.
- Also known as Cross Tabulation table.

Note: When all the values of Contingency table are almost same, we conclude that the design is balanced. Here all values are exactly same, indicating it is a balanced design.

```
# Making Contingency table
table(toothData$supp , toothData$dose)
```

```
    low medium high
  OJ  10     10   10
  VC  10     10   10
```

Now we use boxplot for visualizing the distribution of predictors with respect to the response variable. We will also consider the interaction term in this case. There will be three boxplots, that gives-

- Distribution of tooth length with respect to supplement (supp).
- Distribution of tooth length with respect to dosage (dose).
- Distribution of tooth length with respect to interaction between supp and dose.

Interpretation of first Box Plot: The data that resides in supplement of VC is Normally distributed, whereas the data that resides in OJ is skewed. Not only that the data of OJ completely fits into the data of VC. The Box plot of VC has more variability in it than the Box plot for OJ.

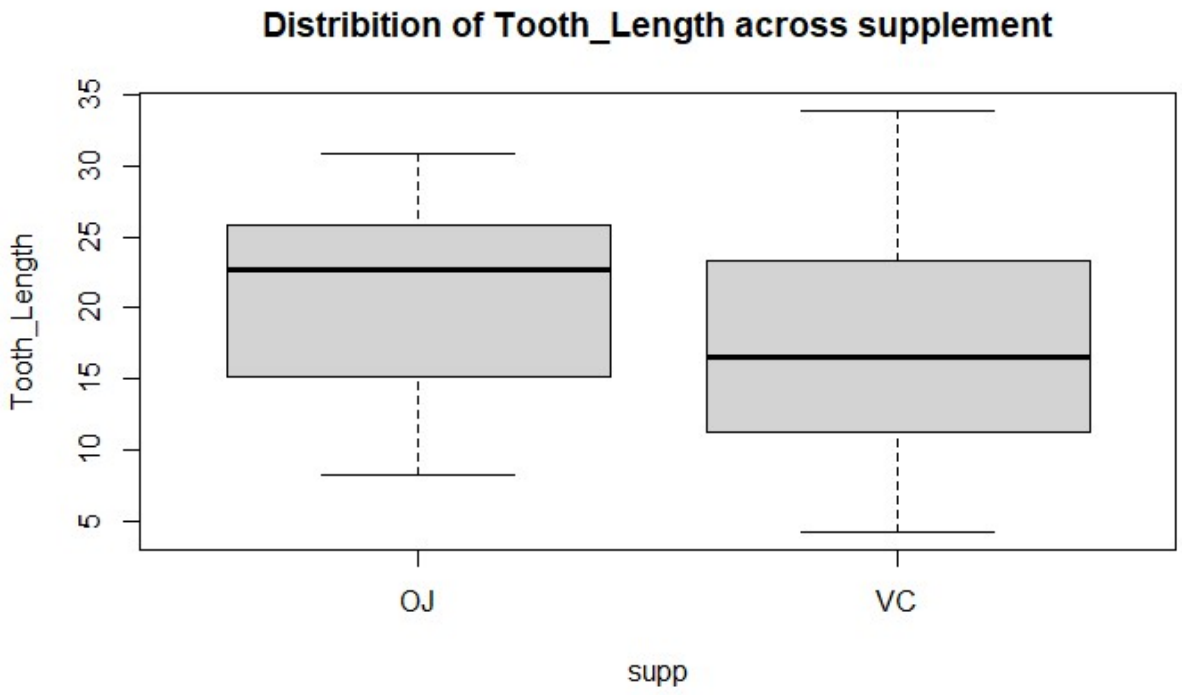
Interpretation of second Box Plot: Here the Box plots do not overlap with each other. All types of doses are normally distributed having same amount of variability. Tooth Length of Guinea Pig that receives high dose of medication is larger than the other doses.

Interpretation of Third Box Plot: Here we consider the distribution on Tooth Length depending on the interaction term of supplement and dose. We observe that

OJ is better way for low dose and VC is better for high dose. VC has more variability and higher value than OJ for high dose.

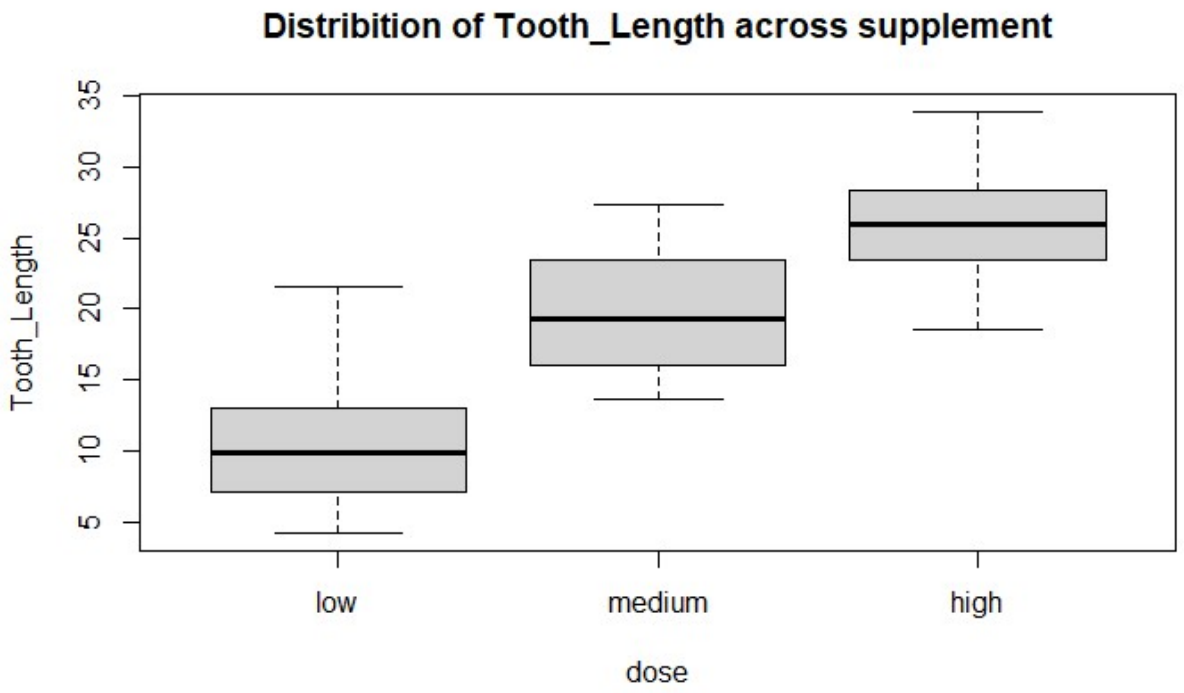
```
# Box plot to show distribution of tooth_length across supp

boxplot(len ~ supp,data= toothData,ylab="Tooth_Length",main="Distribition of Tooth_Length across supplement")
```



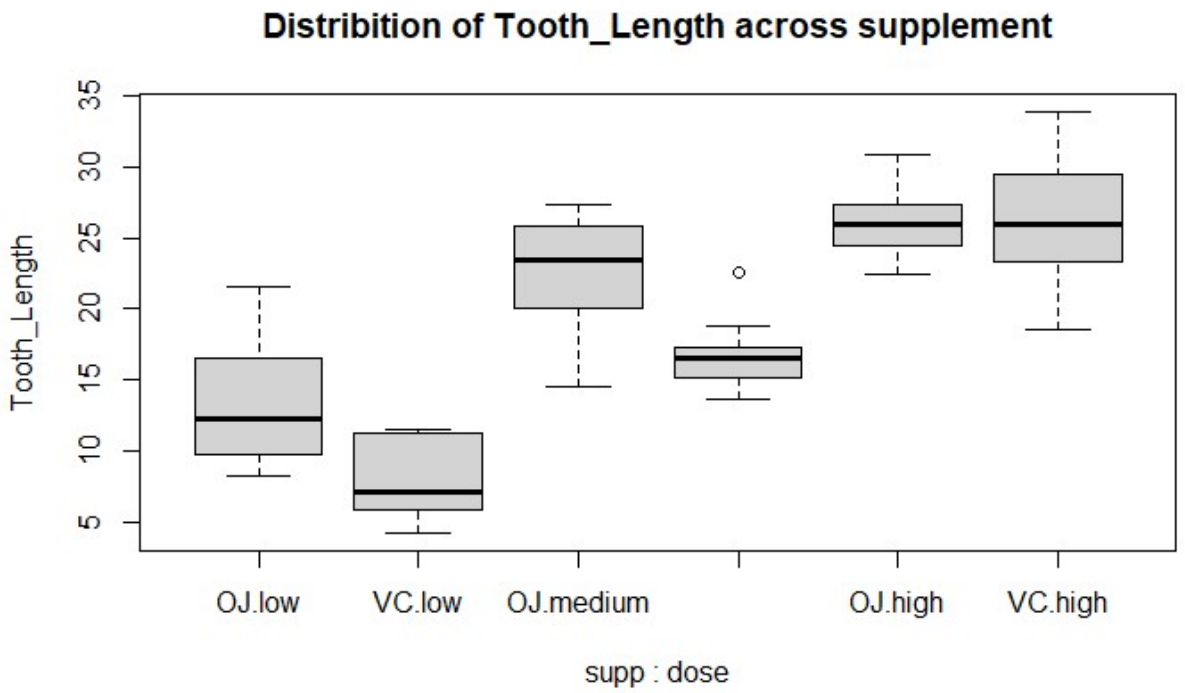

```
# Box plot to show distribution of tooth_length across dose

boxplot(len ~ dose,data= toothData,ylab="Tooth_Length",main="Distribition of Tooth_Length across supplement")
```




```
# Box plot to show distribution of tooth_length across supp and dose interaction effe
ct

boxplot(len ~ supp:dose , data= toothData,ylab="Tooth_Length",main="Distribition of T
ooth_Length across supplement")
```



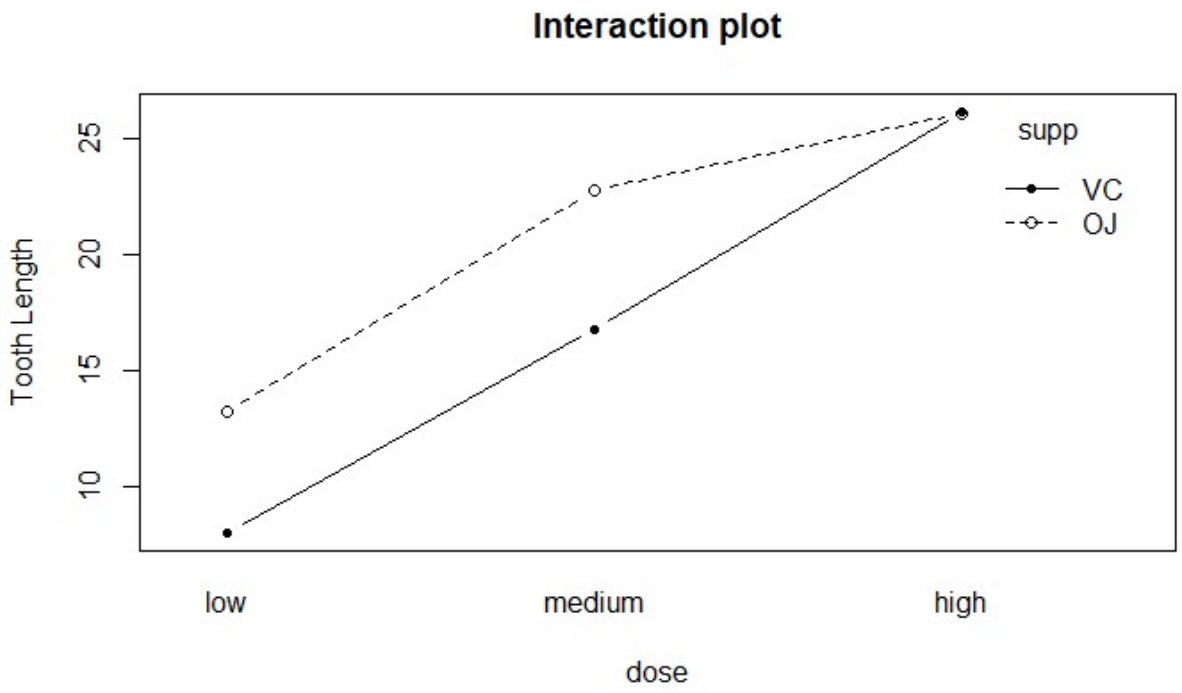
To understand that if there exists any type of interaction effect of supp and dose on response variable, we plot the interaction that.

- This plot is known as interaction plot.
- It concludes on the mean value of Tooth Length.
- Here we use the with() function.

Interpretation of Interaction Plot: This plot gives us some basic ideas such as- * Mean value of Tooth length when supplement VC is at low dose. * Mean value of Tooth length when the supplement VC is at medium dose. * Mean value of Tooth length when the supplement VC is at high dose. * mean value of Tooth length when supplement OJ is at low dose. * mean value of Tooth length when supplement OJ is at medium dose. * mean value of Tooth length when supplement OJ is at high dose.

```
# Plotting interaction Plot for supp and dose variables

with(toothData,interaction.plot(x.factor=dose,trace.factor=supp,response=len,fun=mean
,type="b",legend=T,ylab="Tooth Length",main="Interaction plot",pch=c(1,20)))
```



Now we use tapply function to get the means. It is the points that are visible in the interaction plot.

interpretation of these Numbers: We can interpret the number as the following statements-

- Mean value of Tooth length is 7.98 when supplement VC is at low dose.
- Mean value of Tooth length is 16.77 when the supplement VC is at medium dose.
- Mean value of Tooth length is 26.14 when the supplement VC is at high dose.
- mean value of Tooth length is 13.23 when supplement OJ is at low dose.
- mean value of Tooth length is 22.70 when supplement OJ is at medium dose.
- mean value of Tooth length is 26.06 when supplement OJ is at high dose.

Besides that, we also can conclude some inferences based on the numbers that we get in the below code. These inferences will be-

- Higher the dose, better the tooth growth.
- Gap between the supplements are reducing.

Note: If the interaction plot lines are parallel to each other then we conclude that there is no interaction effect. On the other hand, if the lines meet or tend to meet each other, they we say that there is some interaction effect on the response variable.

```
# In the code below we get the point values that are in the above plot
with(toothData,tapply(len,list(supp,dose),mean))
```

```
     low medium  high
OJ 13.23  22.70 26.06
VC  7.98  16.77 26.14
```

Now we move to the next segment which deals with conduction of the test. One thing to remember is there are mainly two assumptions while conducting the factorial ANOVA in R. They are-

- All the factors should be approximately Normally distributed.
- There must be homogeneity in variance of the factors.

So now our aim is to check for Normality in this data set. We will use Shapiro test to confirm Normality. We use the **shapiro. Test()** function to check Normality. If the P-value, we get from this test is greater than 0.05 (Critical Level) then the data is Normally distributed. Basically, the Null Hypothesis is the underlying data comes from a Normally distributed population.

Interpretation Of Shapiro test: The response variable we consider here is the tooth length. This seems to be **Normally distributed** as the P-value we get is 0.1091, which is higher than typical critical value alpha=0.05.

```
# Conducting Shapiro test on response variable length of tooth for checking Normality
print(shapiro.test(toothData$len))
```

```
    Shapiro-Wilk normality test


data:  toothData$len
W = 0.96743, p-value = 0.1091
```

Now we check the respective variances for the interaction terms in order to conclude the equality of variances. This is given in the code below-

```
# Checking for Homogeneity of variance
with(toothData,tapply(len,list(supp,dose),var))
```

```
      low    medium      high
```

```
OJ 19.889 15.295556  7.049333
VC  7.544  6.326778 23.018222
```

Now by using the below code we are not able to conclude if the Variances are Statistically significantly different or not. To conclude this, we perform **Bartlett test** in R using the same information.

The function we use for this test is **bartlett.test()**. It is an in-built function in R. It takes parameters such as response variable and the interaction terms. If the P-value we get from this test is greater than the critical value alpha=0.05, then we conclude that the variances are Statistically significantly same. In other words, there is Homogeneity in variances.

Interpretation of Bartlett Test: The P-value we get from this test is 0.2261 which is greater than the critical value alpha=0.05. Indicates that there is Homogeneity in variances.

Note: The Bartlett test can also be used to check the Homogeneity in variances between other factors taking single each. This type of examples is given below-

So, conducting Bartlett test on each factor that can affect the response variable we conclude that there is Homogeneity of Variances in the process we consider.

```
# Conducting Bartlett test of Homogeneity of Variances
bartlett.test(len~ interaction(supp,dose),data=toothData)
```

```
    Bartlett test of homogeneity of variances


data:  len by interaction(supp, dose)
Bartlett's K-squared = 6.9273, df = 5, p-value = 0.2261
```

```
# Bartlett test for length of tooth and supplement provided
bartlett.test(len~as.factor(supp),data=toothData)
```

```
    Bartlett test of homogeneity of variances


data:  len by as.factor(supp)
Bartlett's K-squared = 1.4217, df = 1, p-value = 0.2331
```

```
# Bartlett test for length of tooth and doses provided
bartlett.test(len~as.factor(dose),data=toothData)
```

```
    Bartlett test of homogeneity of variances


data:  len by as.factor(dose)
Bartlett's K-squared = 0.66547, df = 2, p-value = 0.717
```

Now we build the model and using factors supp, dose and their interaction term and conduct the factorial ANOVA test.

The **car package** is the mother package of this **Anova()** function. As a parameter it takes the model to be fitted with its type. Here we consider the interaction effect. so, for us the type will be 3. This is given by the below code-

```
# Conducting Factorial ANOVA Test using fitted model and Anova function
attach(toothData)
```

```
The following objects are masked from toothData (pos = 3):


    dose, len, supp


The following objects are masked from toothData (pos = 4):


    dose, len, supp
```

```
fac_aov = Anova(lm(len ~ supp + dose + supp:dose), data=toothData, type=3)
fac_aov
```

```
Anova Table (Type III tests)


Response: len
             Sum Sq Df F value    Pr(>F)
(Intercept) 1750.33  1 132.730 3.603e-16 ***
supp         137.81  1  10.450  0.002092 **
dose         885.26  2  33.565 3.363e-10 ***
supp:dose    108.32  2   4.107  0.021860 *
```

```
Residuals    712.11 54
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Now many times we are disgusted to see the stars in the last column of the above table. So, to omit that we can use the below code-

Interpretation of fac_aov: The mean value of tooth length is Statistically significantly different across 2 levels of supplement and 3 levels of dose. Also, there is significant interaction level.

```
# Ommiting stars from result
options(show.signif.stars=F)
fac_aov
```

```
Anova Table (Type III tests)


Response: len
             Sum Sq Df F value    Pr(>F)
(Intercept) 1750.33  1 132.730 3.603e-16
supp         137.81  1  10.450  0.002092
dose         885.26  2  33.565 3.363e-10
supp:dose    108.32  2   4.107  0.021860
Residuals    712.11 54
```

Now to know the comparison between the three levels of dose, we calculate the marginal values using the emmeans() function, that comes under emmeans package.

It takes the fitted model as a parameter along with the parameter of our interest.

```
# First we create marginal value table for dose variable
fit_model <- lm(len~ supp+dose+supp:dose,data=toothData)
marginal_1 <- emmeans(fit_model , ~dose)
```

```
NOTE: Results may be misleading due to involvement in interactions
```

```
pairs(marginal_1 , adjust = "tukey")
```

```
 contrast      estimate   SE df t.ratio p.value
 low - medium     -9.13 1.15 54  -7.951  <.0001
 low - high      -15.49 1.15 54 -13.493  <.0001
 medium - high    -6.37 1.15 54  -5.543  <.0001

Results are averaged over the levels of: supp
P value adjustment: tukey method for comparing a family of 3 estimates
```

```
# First we create marginal value table for dose variable
fit_model <- lm(len~ supp+dose+supp:dose,data=toothData)
marginal_2 <- emmeans(fit_model , ~supp)
NOTE: Results may be misleading due to involvement in interactions
```

```
pairs(marginal_1 , adjust = "tukey")
 contrast      estimate   SE df t.ratio p.value
 low - medium     -9.13 1.15 54  -7.951  <.0001
 low - high      -15.49 1.15 54 -13.493  <.0001
 medium - high    -6.37 1.15 54  -5.543  <.0001

Results are averaged over the levels of: supp
P value adjustment: tukey method for comparing a family of 3 estimates
# First we create marginal value table for dose variable
fit_model <- lm(len~ supp+dose+supp:dose,data=toothData)
marginal_1 <- emmeans(fit_model , ~ supp:dose)
pairs(marginal_1 , adjust = "tukey")
 contrast              estimate   SE df t.ratio p.value
 OJ low - VC low           5.25 1.62 54   3.233  0.0243
 OJ low - OJ medium       -9.47 1.62 54  -5.831  <.0001
 OJ low - VC medium       -3.54 1.62 54  -2.180  0.2640
 OJ low - OJ high        -12.83 1.62 54  -7.900  <.0001
 OJ low - VC high        -12.91 1.62 54  -7.949  <.0001
 VC low - OJ medium      -14.72 1.62 54  -9.064  <.0001
 VC low - VC medium       -8.79 1.62 54  -5.413  <.0001
```

```
 VC low - OJ high        -18.08 1.62 54 -11.133  <.0001
 VC low - VC high        -18.16 1.62 54 -11.182  <.0001
 OJ medium - VC medium     5.93 1.62 54   3.651  0.0074
 OJ medium - OJ high      -3.36 1.62 54  -2.069  0.3187
 OJ medium - VC high      -3.44 1.62 54  -2.118  0.2936
 VC medium - OJ high      -9.29 1.62 54  -5.720  <.0001
 VC medium - VC high      -9.37 1.62 54  -5.770  <.0001
 OJ high - VC high        -0.08 1.62 54  -0.049  1.0000


P value adjustment: tukey method for comparing a family of 6 estimates
```

Interpretation of the above code: In this case we select the marginal value table for the interaction effect with respect to the response variable length of tooth through the fitted model. It shows that for the below combinations P-value is greater than 0.05 indicating that between them there is no significant difference. The combinations that has no significant difference are-

- OJ high to VC high with P-value 1.0000
- OJ medium to VC high with P-value 0.2936
- OJ medium to OJ high with P-value 0.3187
- OJ low to VC medium with P-value 0.2640