Statistical Inference - Data Analysis

Prasanna Nandakumar

25-Oct-2014

Required:

Analyze the ToothGrowth data in the R datasets package.

Approach:

First we will load/transform the data so that it is close to being normally distributed. We will then find the confidence interval for the mean and then perform a significance test to evaluate whether or not the data is away from a fixed standard. Finally, we will find the power of the test to detect a fixed difference from that standard.

We will assume that a confidence level of 95% is used throughout.

1: Load the ToothGrowth data and perform some basic exploratory data analyses

```r
require(graphics)

#install.packages("UsingR")
#library(UsingR)

data(ToothGrowth)
tgrowth <- ToothGrowth$len
tsupp <- ToothGrowth$supp
tdose <- ToothGrowth$dose
tgdata<- ToothGrowth
head(ToothGrowth,2)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
```

```r
tail(ToothGrowth,2)
```

```
##     len supp dose
## 59 29.4   OJ    2
## 60 23.0   OJ    2
```

Basic exploratory data analyses (BEDA) There are no rownames for the data; each line is numbered starting from 1. There are three columns in the dataframe: "len," "supp," and "dose."

```r
dim(ToothGrowth)
```

```
## [1] 60  3
```

```
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 ...
```

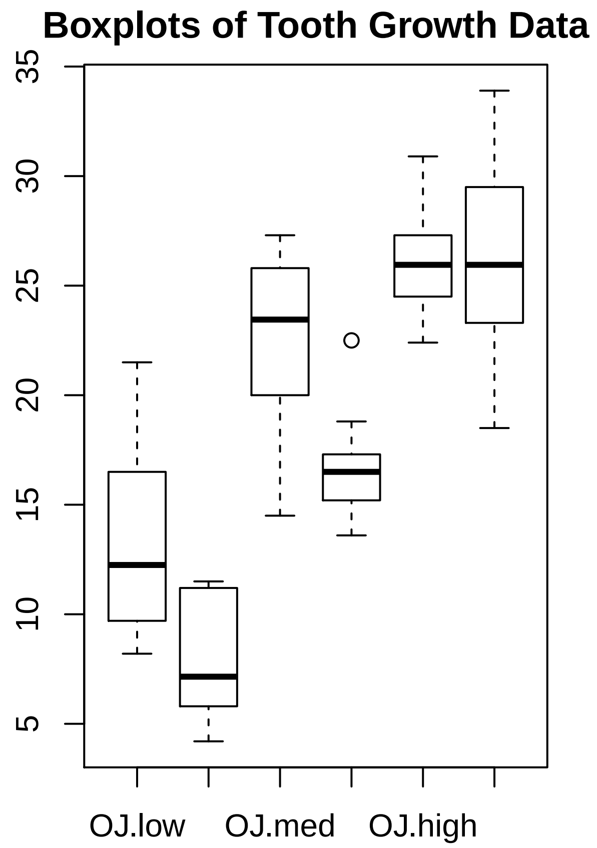
The data set is contains 30 observations for each of the supplements-Orange Juice (OJ) and Vitamin C (VC). There are no Null or NA values.

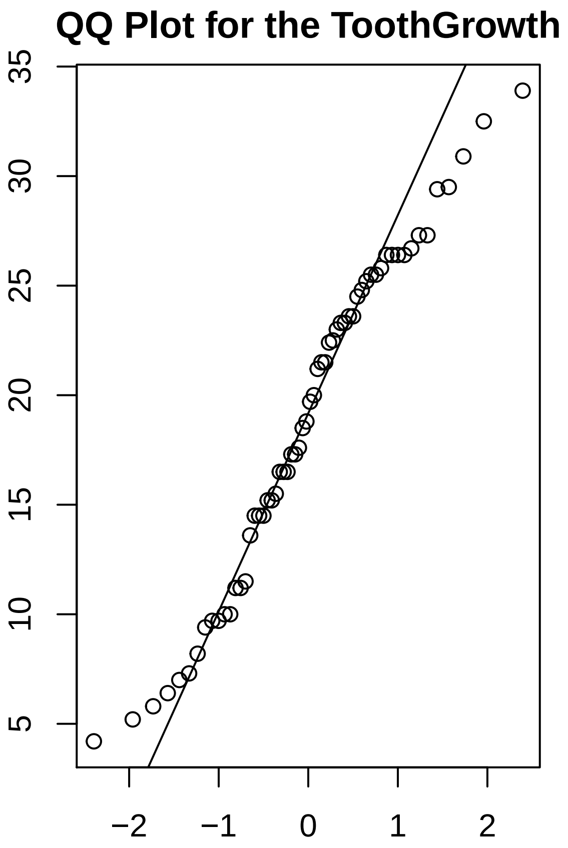
I replaced the dose levels with low, med, high for 0.5, 1.0, 2.0 milligrams. We subset data into two dataframes by supplement type.
For each row: len: numeric variable giving the length of teeth. supp: categorical variable (factor) representing the supplement:“VC” (vitamin C) or “OJ” (orange juice). dose is a numeric variable representing the amount of supplement in milligrams.

```
ToothGrowth$dose = factor(ToothGrowth$dose, levels=c(0.5,1.0,2.0),labels=c("low","med","high"))
tgOJ <- ToothGrowth[ToothGrowth$supp == 'OJ',]
tgVC <- ToothGrowth[ToothGrowth$supp == 'VC',]
# set up a two panel plot
par(mfrow=c(1,2))
par(mar=c(2,2,2,2)+0.1)
#plot(ToothGrowth)


boxplot(len ~ supp * dose, data=ToothGrowth,
 ylab="Tooth Length", main="Boxplots of Tooth Growth Data")


qqnorm(ToothGrowth$len,main="QQ Plot for the ToothGrowth")
qqline(ToothGrowth$len)
```





```r
bartlett.test(ToothGrowth$len~supp,ToothGrowth)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  ToothGrowth$len by supp
## Bartlett's K-squared = 1.4217, df = 1, p-value = 0.2331
```

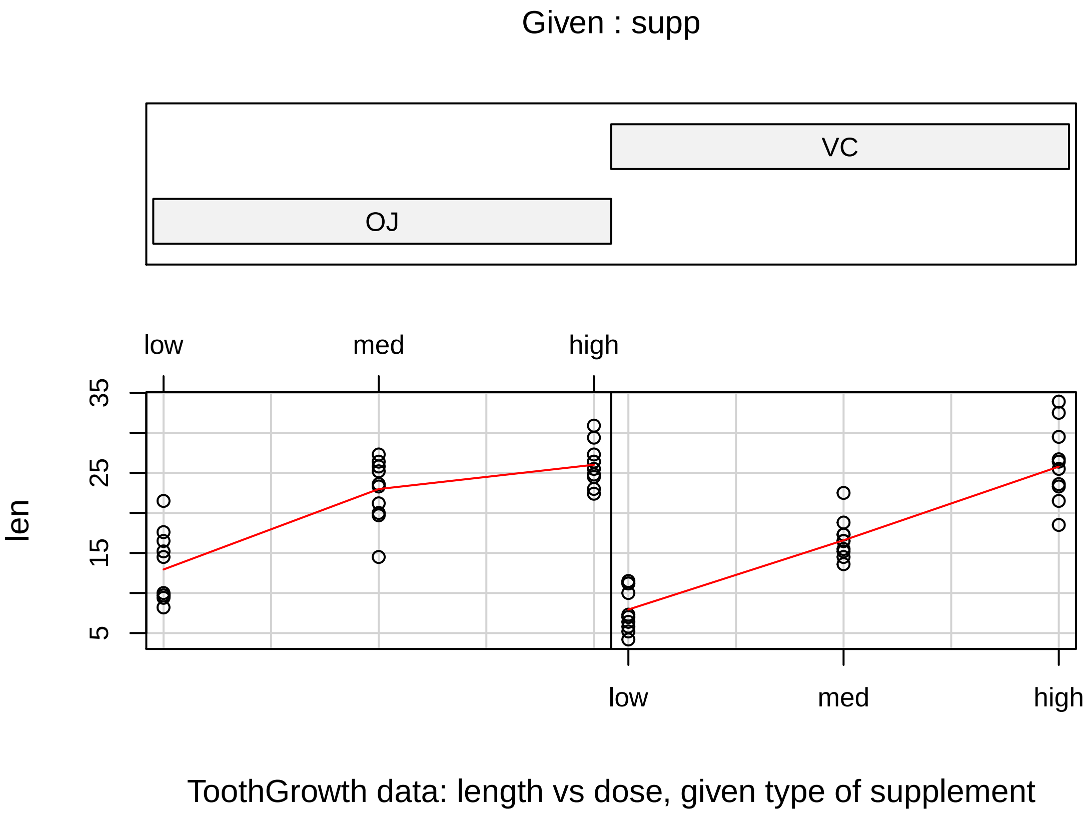
```r
bartlett.test(ToothGrowth$len~dose,ToothGrowth)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  ToothGrowth$len by dose
## Bartlett's K-squared = 0.6655, df = 2, p-value = 0.717
```

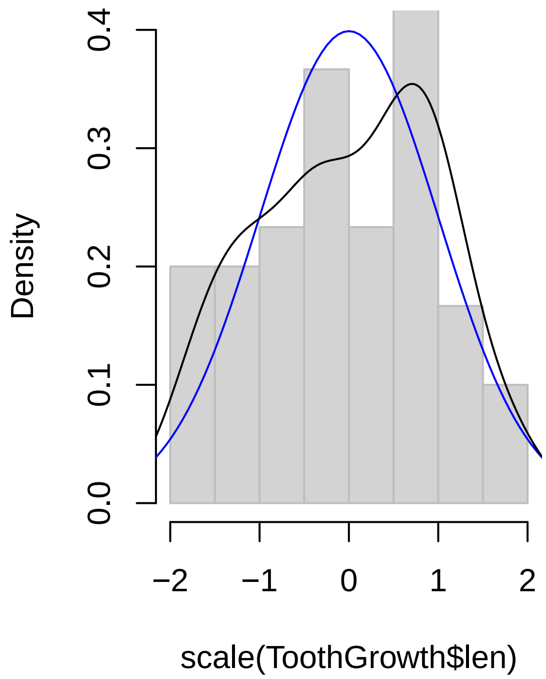
From the box-plots, two things are apparent. - First, the data looks relatively normal, there are no signs of severe skewness or the presence of outliers in any. - Second, it looks pretty clear that increases in the dosage of vitamin C and Orange Juice given to the test subjects leads to increases in tooth length. Their is one point outlier in medium dose group in VC supplement, rest of the data is arranged in distinct groups.

QQ graph shows ToothGrowth is approximately normally distributed with some deviation at the especially near the extreme ends of the normal distribution quantiles.

```
## Commenting all this section so I fit into 2 pages as required; very tough job.
par(mfrow=c(1,2))
coplot(len ~ dose | supp, data = ToothGrowth, panel = panel.smooth,
 xlab = "ToothGrowth data: length vs dose, given type of supplement")
```



```
#
#
hist(scale(ToothGrowth$len), prob=T, col="light grey", border="grey", main=NULL, ylim=c(0,0.4))
curve(dnorm(x,0,1), -3, 3, col='blue', add=T)
lines(density(scale(ToothGrowth$len)))
#
#with(ToothGrowth, tapply(len, list(supp,dose), mean))
```



2: Provide a basic summary of the data.

```
summary(ToothGrowth)
```

```
##       len         supp      dose   
##  Min.   : 4.20   OJ:30   low :20  
##  1st Qu.:13.07   VC:30   med :20  
##  Median :19.25           high:20  
##  Mean   :18.81                    
##  3rd Qu.:25.27                    
##  Max.   :33.90                    
```

```
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: Factor w/ 3 levels "low","med","high": 1 1 1 1 1 1 1 1 1 1 ...
```

```
x1 <- mean(tgOJ$len)
x2 <- mean(tgVC$len)
tgmean <- mean(ToothGrowth$len)
```

```r
s1 <- sd(tgOJ$len)
s2 <- sd(tgVC$len)
sdtg <- sd(ToothGrowth$len)
n1 <- length(tgOJ$len)
n2 <- length(tgVC$len)
ntg <- length(ToothGrowth$len)
SE <- sqrt( s1^2/n1 + s2^2/n2)
tgse <- (sdtg/sqrt(ntg))
tgerr <- tgse*qt(0.975,df=ntg-1)
tgleft <- tgmean-tgerr
tgright <- tgmean+tgerr

tgCI <- (mean(ToothGrowth$len) + c(-1, 1) * qnorm(0.975) * sd(ToothGrowth$len)/sqrt(length(ToothGrowth$l
### commenting this to fit in 2 pages.
### Typing the reuslts to save on space.
cat("The Mean of ToothGrowth with supp OJ =",x1,"The Mean of ToothGrowth with supp VC =",x2,"\n")
```

```
## The Mean of ToothGrowth with supp OJ = 20.66333 The Mean of ToothGrowth with supp VC = 16.96333
```

```r
cat("The sd of ToothGrowth with supp OJ =",s1,"The sd of ToothGrowth with supp VC =",s2,"\n")
```

```
## The sd of ToothGrowth with supp OJ = 6.605561 The sd of ToothGrowth with supp VC = 8.266029
```

```r
cat("The length of dataframe wiht supp OJ =",n1,"The length of dataframe wiht supp VC =",n2,"\n")
```

```
## The length of dataframe wiht supp OJ = 30 The length of dataframe wiht supp VC = 30
```

```r
cat("The Standard Error for OJ and VC=",SE,"\n")
```

```
## The Standard Error for OJ and VC= 1.931844
```

```r
cat("The 95% CI for the Tooth Growth is between =",tgCI,"\n")
```

```
## The 95% CI for the Tooth Growth is between = 16.87783 20.74884
```

```r
cat("The Mean for TothGrowth for OJ and VC=",mean(ToothGrowth$len),"\n")
```

```
## The Mean for TothGrowth for OJ and VC= 18.81333
```

```r
#comment out to fit into 2 pages
shapiro.test(ToothGrowth$len)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ToothGrowth$len
## W = 0.9674, p-value = 0.1091
```

- The Mean of ToothGrowth with supp OJ = 20.66 and the Mean of ToothGrowth with supp VC = 16.96.
- The sd of ToothGrowth with supp OJ = 6.606, and the sd of ToothGrowth with supp VC = 8.266.
- The length of dataframe with supp OJ = 30, and the length of dataframe with supp VC = 30.
- The Standard Error for OJ and VC= 1.932, and the 95% CI for the Tooth Growth is between = 16.88 20.75.
- The Mean for TothGrowth for OJ and VC= 18.81.

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

(Use the techniques from class even if there's other approaches worth considering) The 95% CI for the Tooth Growth is between = 16.88 20.75. The margin of error is found based on a 95% confidence level is 1.976028. Hence the 95% confidence interval for the tooth growth is between 16.88 and 20.75.

```
# Comment out to fit into pages.
t <- (x1 - x2)/SE
px<-pt(t,n1-1)
#cat("t Vlaue is=",t,"P vlaues is =",px)
#sdf <- tgdata[sample(nrow(tgdata), 40), ]
#smean <- mean(sdf$len)
#smean
summary(aov(len~supp+dose,data=ToothGrowth))
```

```
##             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
```

```
summary(aov(len~supp+dose+supp:dose,data=ToothGrowth))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## supp         1  205.4   205.4  15.572 0.000231 ***
## dose         2 2426.4  1213.2  92.000  < 2e-16 ***
## supp:dose    2  108.3    54.2   4.107 0.021860 *  
## Residuals   54  712.1    13.2                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(len~supp*dose,data=ToothGrowth))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## supp         1  205.4   205.4  15.572 0.000231 ***
## dose         2 2426.4  1213.2  92.000  < 2e-16 ***
## supp:dose    2  108.3    54.2   4.107 0.021860 *  
## Residuals   54  712.1    13.2                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(len~supp*dose,data=ToothGrowth, subset=(dose!="2")))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## supp         1  205.4   205.4  15.572 0.000231 ***
## dose         2 2426.4  1213.2  92.000  < 2e-16 ***
## supp:dose    2  108.3    54.2   4.107 0.021860 *  
## Residuals   54  712.1    13.2                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#subset by dose
big_dose <- ToothGrowth[ToothGrowth$dose==2.0,]
med_dose <- ToothGrowth[ToothGrowth$dose==1.0,]
small_dose <- ToothGrowth[ToothGrowth$dose==0.5,]

# mean by dose
bigd_mean <- mean(big_dose$len)
medd_mean <- mean(med_dose$len)
smalld_mean <- mean(small_dose$len)
# sd by dose
bigd_sd <- sd(big_dose$len)
medd_sd <- sd(med_dose$len)
smalld_sd <- sd(small_dose$len)
# df by dose
bigd_n <- length(big_dose$len)
medd_n <- length(med_dose$len)
smalld_n <- length(small_dose$len)
does_n <- bigd_n+medd_n+smalld_n
#standard erro by dose
dose_SE <- sqrt( bigd_sd^2/bigd_n + medd_sd^2/medd_n + smalld_sd^2/smalld_n)
dose_err <- dose_SE*qt(0.975,df=does_n-1)
```

```
## Warning in qt(0.975, df = does_n - 1): NaNs produced
```

```
# Crtical t-value for big and med dose
t_bm <- (bigd_mean - medd_mean)/dose_SE
# Crtical t-value for med and small dose
t_ms <- (medd_mean - smalld_mean)/dose_SE
# P-Value for big and med dose
p_bm <- pt(t_bm,does_n - 1 )
# P-Value for med and small dose
p_ms <- pt(t_ms,does_n - 1 )

t.test(tgOJ$len,tgVC$len)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  tgOJ$len and tgVC$len
## 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 of x mean of y
##  20.66333  16.96333
```

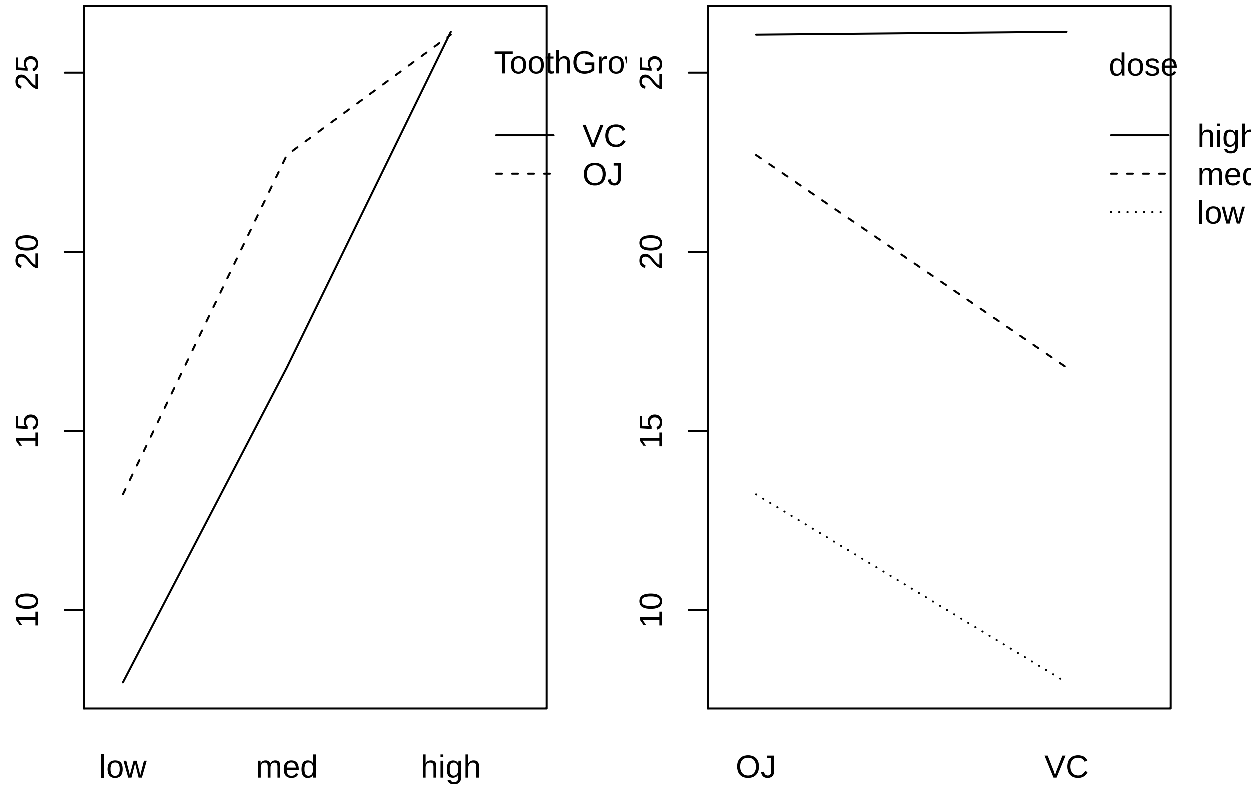
Ho:meu_OJ = meu_VC Ha:meu_OJ != Meu_VC our hypothesis is that Orange Juice and Vitamin see have the similar effect on the totth growth. We got t-vlaue of 1.915268 and p of 0.9673165, we reject the Null hypothesis.

Thus there is different effect with respect to the supplements. Next we will evaluate, lower dose levels effect is similar to higher dose levels by evlauting the Null Hypothesis: H0: meu_hd = mue_ld Ha: meu_hd != mue_ld We got t-vlaue of 3.87 and p-value of 0.9999 for higher doses and t-value 5.55 and p-value 0.9999, we reject Null Hypothesis i.e. lower doses have more effect on the tooth growth than higher doeses. From graphs we can see that effect at higher dose (2.0) is almost same for the OJ and VC.

4: State your conclusions and the assumptions needed for your conclusions.

```
# Comment out to fit into pages.
par(mfrow=c(1,2))
par(mar=c(2,2,2,2)+0.1)
interaction.plot(ToothGrowth$dose, ToothGrowth$supp, ToothGrowth$len)

#interaction plot
with(ToothGrowth, {interaction.plot(supp, dose, len)})
```



```
summary(aov(len~supp*dose,data=ToothGrowth, subset=(dose!="2")))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## supp         1  205.4   205.4  15.572 0.000231 ***
## dose         2 2426.4  1213.2  92.000  < 2e-16 ***
## supp:dose    2  108.3    54.2   4.107 0.021860 *  
## Residuals   54  712.1    13.2                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the statistical analysis and graphs it is pretty clear that increases in the dosage of vitamin C given to the subjects leads to increases in tooth length and the same is true with orange juice.

We can see in the plots and , that the tooth length increases as dosage increases, and at two low dosages, OJ works better than VC with similar effects, but at the high dosage, OJ and VC shows no difference.

The effect of supplement does depend on levels of dose, therefore, there is a supp-dose interactions. Although, there is no placebo data hence we can't realy determine the real advantage of each supplement with respect no supplement, but still the graphs and statistics indicate there is significant effect especially at lowe doses with vitamin c and orange juice on tooth growth of the subjects.