Statistical inference Course Project part2 with exploratory analysis

Author : Aditya Gudal

INSTRUCTIONS

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)

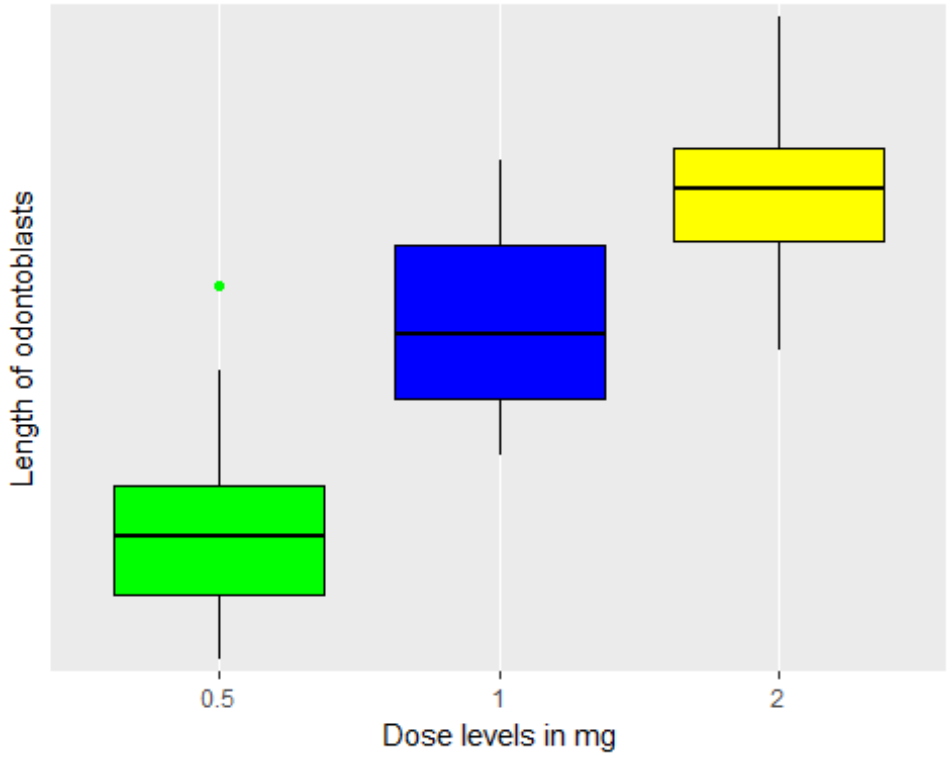
4.State your conclusions and the assumptions needed for your conclusions.

OBSERVATIONS AND INFERENCE

Here the dataset is taken from the datasets package in R .Here we have the response being the length of odonoblasts in 60 guinea pigs.Each animal received one of the three dose levels of viatmain C(0.5,1,2mg/day) by on of the two supplements that is Ascorbic acid(VC) and Orange juice(OJ).

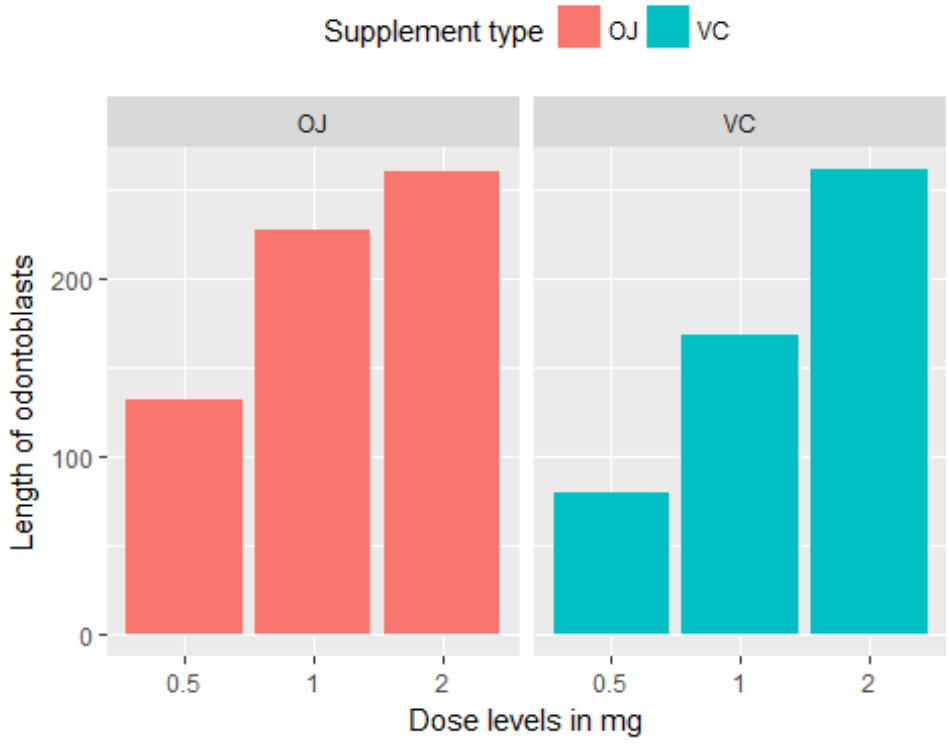
Plotting system used is ggplot2 I have shown the boxplot of the two variables or features taken dose vs length.The plot shows three doses and how theyvary with length .Dose 1 has an outlier as shown.Rest of the doses are stable.

```
library(datasets)
library(ggplot2)
ToothGrowth$dose=as.factor(ToothGrowth$dose)
g<-ggplot(data=ToothGrowth,mapping=aes(x=dose,y=len))
ge<-g+geom_boxplot(color='black',outlier.colour =
'green',fill=c('green','blue','yellow'))
ge+scale_x_discrete(name='Dose levels in mg')+scale_y_discrete(name="Length
of odontoblasts")
```



Here I have plotted again dose vs length but taking into consideration the supplements and how they vary using a bar chart.This shows how the length growth using dose 2 with 2mg per day is similar in effect using both the supplements OJ And VC.

```
p<-ggplot(data=ToothGrowth,mapping=aes(x=dose,y=len,fill=supp))
p+geom_bar(stat='identity')+facet_grid(.~supp)+guides(fill=guide_legend(title
='Supplement type'))+theme(legend.position = 'top')+
  xlab('Dose levels in mg')+
  ylab('Length of odontoblasts')
```



Describes the entire dataset.

```
summary(ToothGrowth)
```

```
##       len        supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
##  Median :19.25           2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

The datasets dimensions.

```
dim(ToothGrowth)
```

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

I have then used a linear model to fit the data using simple linear regression. The following summary gives us a great insight on the data. Shows the adjusted R2 test and the following p values that are below the assumed value 0.05.

```
linear<-lm(formula = len~dose+supp,data=ToothGrowth)
summary(linear)
```

```
##
## Call:
## lm(formula = len ~ dose + supp, data = ToothGrowth)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.085 -2.751 -0.800  2.446  9.650
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.4550     0.9883  12.603  < 2e-16 ***
## dose1         9.1300     1.2104   7.543 4.38e-10 ***
## dose2        15.4950     1.2104  12.802  < 2e-16 ***
## suppVC       -3.7000     0.9883  -3.744 0.000429 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.828 on 56 degrees of freedom
## Multiple R-squared:  0.7623, Adjusted R-squared:  0.7496
## F-statistic: 59.88 on 3 and 56 DF,  p-value: < 2.2e-16
```

The 95% confidence interval as shown.

```
confint(linear,level=0.95)
```

```
##                 2.5 %    97.5 %
## (Intercept) 10.475238 14.434762
## dose1        6.705297 11.554703
## dose2       13.070297 17.919703
## suppVC      -5.679762 -1.720238
```

With low p values and confidence interval we can reject the null hypothesis that the coeffients or weights are zeroes .In addition we see that t values of each variable are in the range except dose 2 that has its t value being slightly lower than its confidence interval.