ToothGrowth analysis

Unmesh Phadke

24th July 2015

Description

This report is part of the course project for the course Statistical Inference which is part of the Data Science Specialisation offered by Johns Hopkins University on Coursera. This project deals with the ToothGrowth dataset in R.In this dataset,the response is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid). The dataframe has 60 observations with the following 3 variables: 1)len :numeric Tooth Length 2)supp factor: Supplement type(VC or OJ) 3)dose: numeric Dose in kilograms

The main goals for this project are: 1.To perform basic exploratory data analyses of the ToothGrowth data set in R. 2.To summarise the dataset. 3.Compare tooth-growths by suplement and dosage using hypothesis testing and/or confidence intervals.

Exploratory Data Analyses

Now we will do some basic exploratory data analysis on the data set.

First,we will convert the ‘supp’ variable in the dataset to a factor variable.

```
data(ToothGrowth)
ToothGrowth$supp<-as.character(as.factor(ToothGrowth$supp))
```

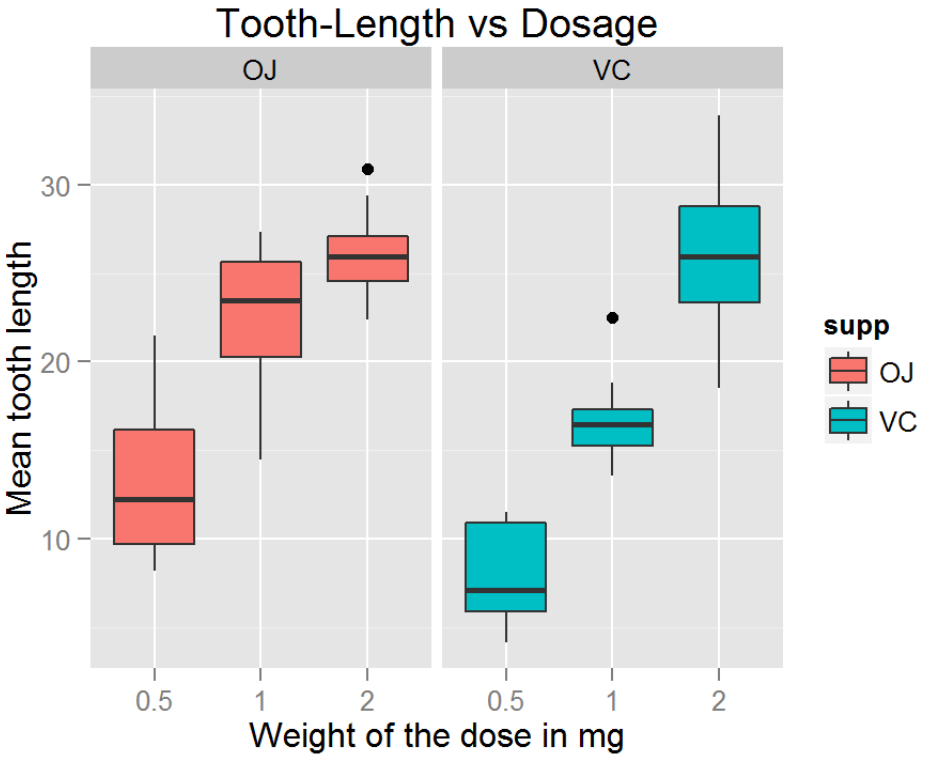
Now we will plot the data. For each type of delivery method (OJ and VC) we will plot the mean tooth growth for each dosage.

```
graph<-ggplot(df,aes(df$dose,df$len))
graph<-graph+ geom_point()
graph<-graph+ geom_line()
graph<- graph + facet_grid(Group.2~.) #Facets for the 'supp' variable
graph<-graph + labs(x= "Weight of the dose in mg")+ labs(y= "Mean tooth length") + lab
s(title= "Tooth-Length vs Dosage")
graph
```


```
remove(graph)
```

Drawing boxplots could be useful as well. We do this below.

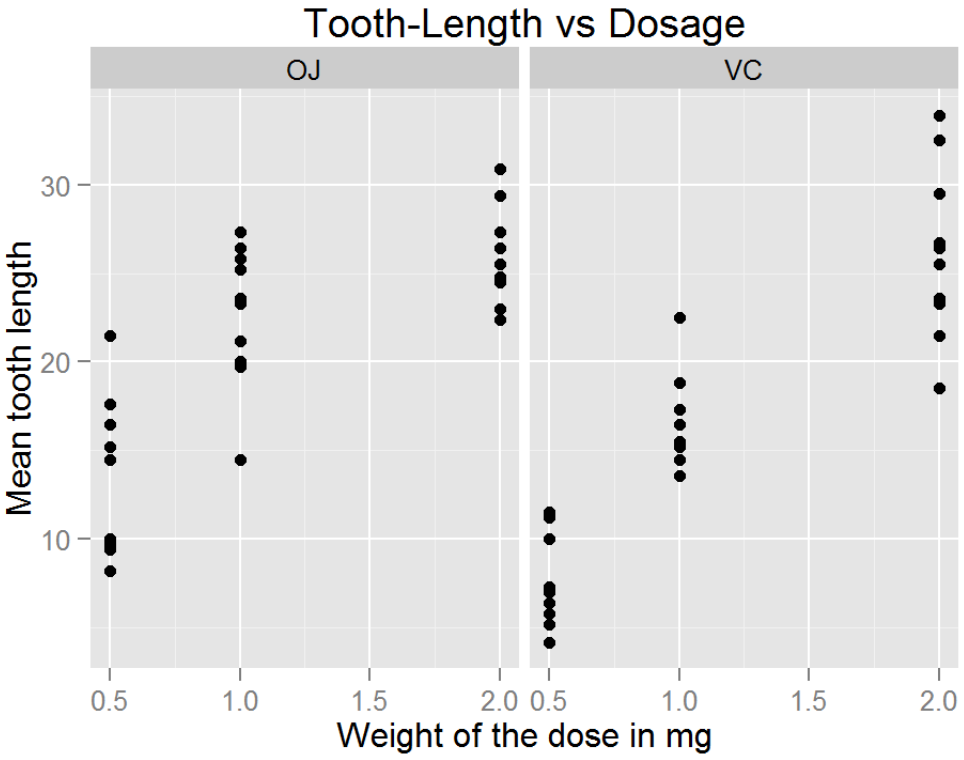
```
g<-ggplot(ToothGrowth,aes(factor(dose),len,fill=supp))
g<-g + geom_boxplot()
g<-g + facet_grid(.~supp)
g<-g +labs(x= "Weight of the dose in mg")+ labs(y= "Mean tooth length") + labs(title=
"Tooth-Length vs Dosage")
g
```




```
remove(g)
```

We would also like to probably look at the variance for each dosage weight . Simple scatter plots do the job well here.

```
g1<-ggplot(ToothGrowth,aes(dose,len)) #Note that the dataframe used is #ToothGrowth and
not df.
g1<-g1 + geom_point()
g1<-g1+ facet_grid(.~supp)
g1<-g1 +labs(x= "Weight of the dose in mg")+ labs(y= "Mean tooth length") + labs(title=
"Tooth-Length vs Dosage")
g1
```




```
remove(g1)
```

It can be noted that for VC,the variance in tooth size is pretty high for the 2,g dosage..

Short Summary of the data

Here is a short summary of the dataset.

```
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: chr  "VC" "VC" "VC" "VC" ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

```
summary(ToothGrowth)
```

```
##       len            supp                dose
##  Min.   : 4.20   Length:60          Min.   :0.500
##  1st Qu.:13.07   Class :character   1st Qu.:0.500
##  Median :19.25   Mode  :character   Median :1.000
##  Mean   :18.81                      Mean   :1.167
##  3rd Qu.:25.27                      3rd Qu.:2.000
##  Max.   :33.90                      Max.   :2.000
```

```
df_VC<-subset(ToothGrowth,supp=="VC")
df_OJ<-subset(ToothGrowth,supp=="OJ")
tapply(ToothGrowth$len,ToothGrowth$supp,FUN=summary)
```

```
## $OJ
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.20   15.52   22.70   20.66   25.72   30.90
##
## $VC
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.20   11.20   16.50   16.96   23.10   33.90
```

Comparing tooth length with respect to supp and dose

In this section we will perform hypothesis testing to compare dependence of tooth growth on the dosage and the supp variable.

So, here we will perform a paired t test for each dosage value comparing whether the means for the two supp values are equal or not. We first do the analysis for the dosage of 0.5 mg.

```
df_VC<-group_by(df_VC,dose)
df_OJ<-group_by(df_OJ,dose)
t.test(df_OJ$len[1:10],df_VC$len[1:10],paired=TRUE,var.equal=FALSE) #Test for 0.5 mg
```

```
##
##  Paired t-test
##
## data:  df_OJ$len[1:10] and df_VC$len[1:10]
## t = 2.9791, df = 9, p-value = 0.01547
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.263458 9.236542
## sample estimates:
## mean of the differences
##                    5.25
```

From the output we observe that the p-value is 0.015 which is much less than 0.05 and the confidence intervals(95%) also, does not contain 0. So the evidence is enough to reject the null hypothesis. The confidence interval has positive values so the tooth length is expected to be higher in case of the Orange Juice dosage for 0.5 mg.

```
t.test(df_OJ$len[11:20],df_VC$len[11:20],paired=TRUE,var.equal=FALSE) #Test for 1 mg
```

```
##
##  Paired t-test
##
## data:  df_OJ$len[11:20] and df_VC$len[11:20]
## t = 3.3721, df = 9, p-value = 0.008229
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.951911 9.908089
## sample estimates:
## mean of the differences
##                    5.93
```

Again, we observe that the P-value is 0.008 which is very less and this presents strong evidence against the null hypothesis.So, similarly as the previous case, the expected tooth length is greater for the orange juice doasge for the 1 mg case.

```
t.test(df_OJ$len[21:30],df_VC$len[21:30],paired=TRUE,var.equal=FALSE) #Test for 1.5 mg
```

```
##
##  Paired t-test
##
## data:  df_OJ$len[21:30] and df_VC$len[21:30]
## t = -0.042592, df = 9, p-value = 0.967
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.328976  4.168976
## sample estimates:
## mean of the differences
##                   -0.08
```

In this case, it is observed that the P-value is very high at 0.967 which presents strong evidence in favour of the null hypothesis. Also note that the 95% confidence interval has zero contained in this and so the null hypothesis that the expected tooth growths for both dosages pf 2mg is accepted. Note that this is also vaguely evident from the scatter plot and the box-plot in the exploratory analysis section.

In the above section, while performing the simulations we assumed the following things: 1:The true variances of the sample are unequal. This assumption is made because we are not sure about the parameters of the distributions for the two different types of supplies and the factors that affect these distrributions. In sucha case it is safe to assume unequal variances. 2. We are assuming that a threshold of 0.05 for the P-values is enough to accept one hypothesis. That means that we are dealing with 95 % confidence intervals. 3.The data is paired as the corresponding readings for the two supply categories are for the same pig(10 different pigs).

1:The dosage using the Orange Juice is proving to be more effective in general than the one using ascorbic acid atlest for the 0.5mg and 1 mg doses. For the 2 mg dose, both perform equally well. 2:Exploratory data analyses confirm that expected tooth length is dependent on the amount of dose given and whether it is given through Orange juice or Ascorbic acid.