Tooth Growth Data Analysis

Aanish Singla

1 April 2016

Overview: In this document we will analyze the ToothGrowth dataset that comes with R "datasets" and do hypothesis testing for facts which are not obvious.

Lets see variables and the values:

```
library("datasets")
str(ToothGrowth)
```

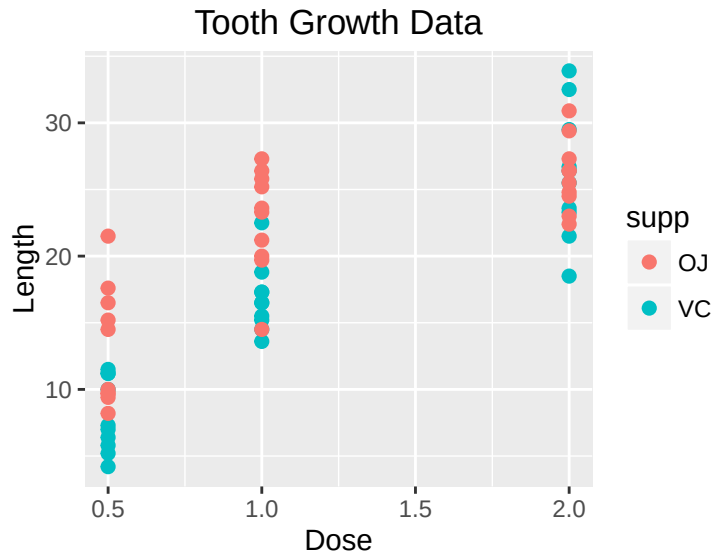
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

There are 3 variables - len, supp and dose. Supp (Delivery Method) is a 2 level factor with value "VC" and "OJ". Other 2 fields are numeric.

Also we can see that there are 60 rows in the dataset.

Lets see how the plot looks for length and dose for different supp

```
library(ggplot2)
g <- ggplot(ToothGrowth, aes(dose, len))
g + geom_point(aes(color=supp), size=2) +
  xlab("Dose") + ylab("Length") + ggtitle("Tooth Growth Data")
```



There is a clear increase in length as the Dose increases.

What about the effect of supp(delivery method) on length?

Lets frame a hypothesis for this. Null Hypothesis: H0 = Delivery method has no effect on length i.e. H0 = Mean length with OJ = Mean length with VC

Alternate Hypothesis = Ha = Delivery method has an effect on length i.e. Ha = Mean length with OJ <> Mean length with VC

Lets perform a two sided non-paired T-test to verify the Hypothesis (assuming variances are not equal)

```
t.test(ToothGrowth$len~ToothGrowth$supp, paired=F,
       var.equal=F, alt="two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len by ToothGrowth$supp
## 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 in group OJ mean in group VC
##         20.66333         16.96333
```

Since the p-value is 0.06063 (greater than 0.05), we fail to reject the Null hypothesis. Also we can notice that that 95% confidence interval contains zero in it, so we have no statistical evidence that one delivery method is better than other.