The *ToothGrowth* data analysis.

Maciej Nowak

Tuesday, July 21, 2015

Exploratory data analysis

We will be working on the *ToothGrowth* data that comes with R. The R documentation titles this data as *"The Effect of Vitamin C on Tooth Growth in Guinea Pigs"* and that Vitamin C was served as orange juice (OJ) or ascorbic acid (VC - "pure" Vitamin C). First we load the data:

```
rm(list = ls()) # clear the environment so the results are reproducible
data(ToothGrowth) # load the data
class(ToothGrowth) # check the class/data type
```

```
## [1] "data.frame"
```

Now that we know it is a *data.frame* we can continue our analysis:

```
head(ToothGrowth, 3) # check the columns and how the data looks like
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
```

```
nrow(ToothGrowth) # how many rows are there?
```

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

```
levels(ToothGrowth$supp) # how many supplement types are there?
```

```
## [1] "OJ" "VC"
```

```
unique(ToothGrowth$dose) # is the dose discrete or continuous?
```

```
## [1] 0.5 1.0 2.0
```

```
ToothGrowth[ which.max(ToothGrowth$len), ] # the record with max len
```

```
##     len supp dose
## 23 33.9   VC    2
```

```
ToothGrowth[ which.min(ToothGrowth$len), ] # the record with min len
```

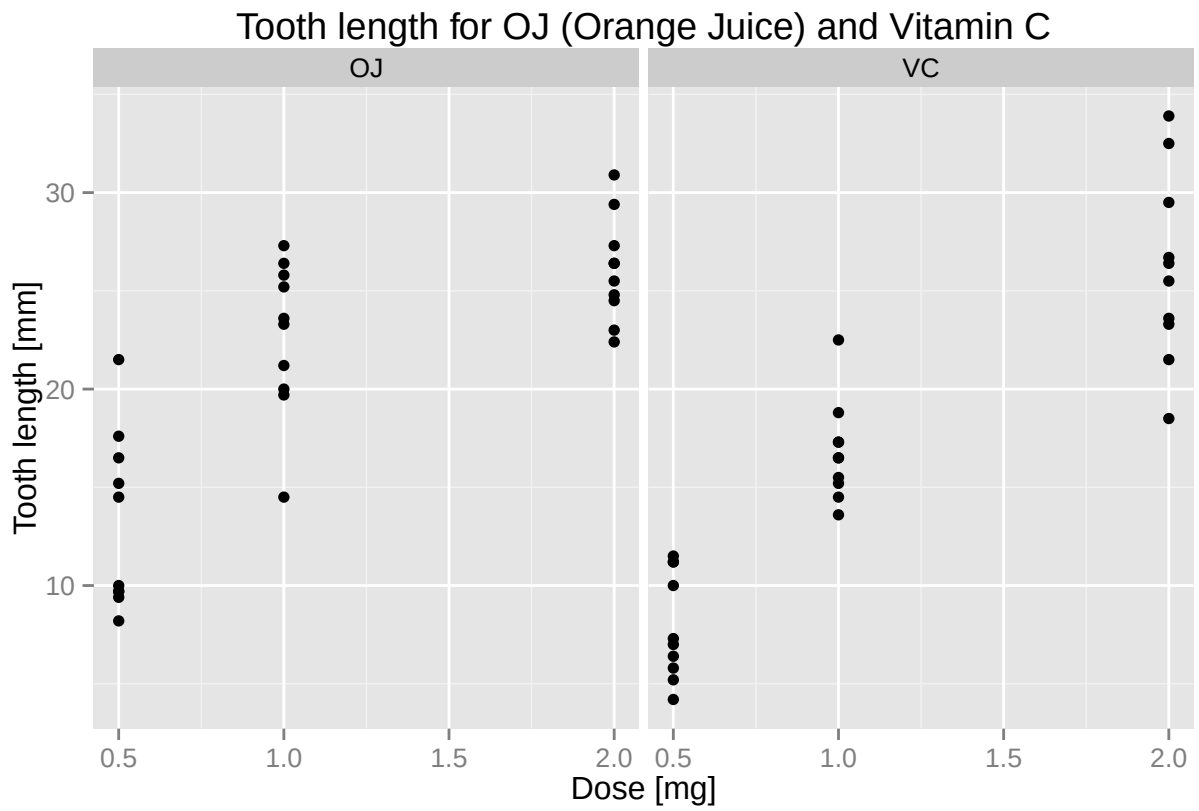
```
##   len supp dose
## 1 4.2   VC  0.5
```

```r
which(is.na(ToothGrowth$len)) # any NA len
```

```
## integer(0)
```

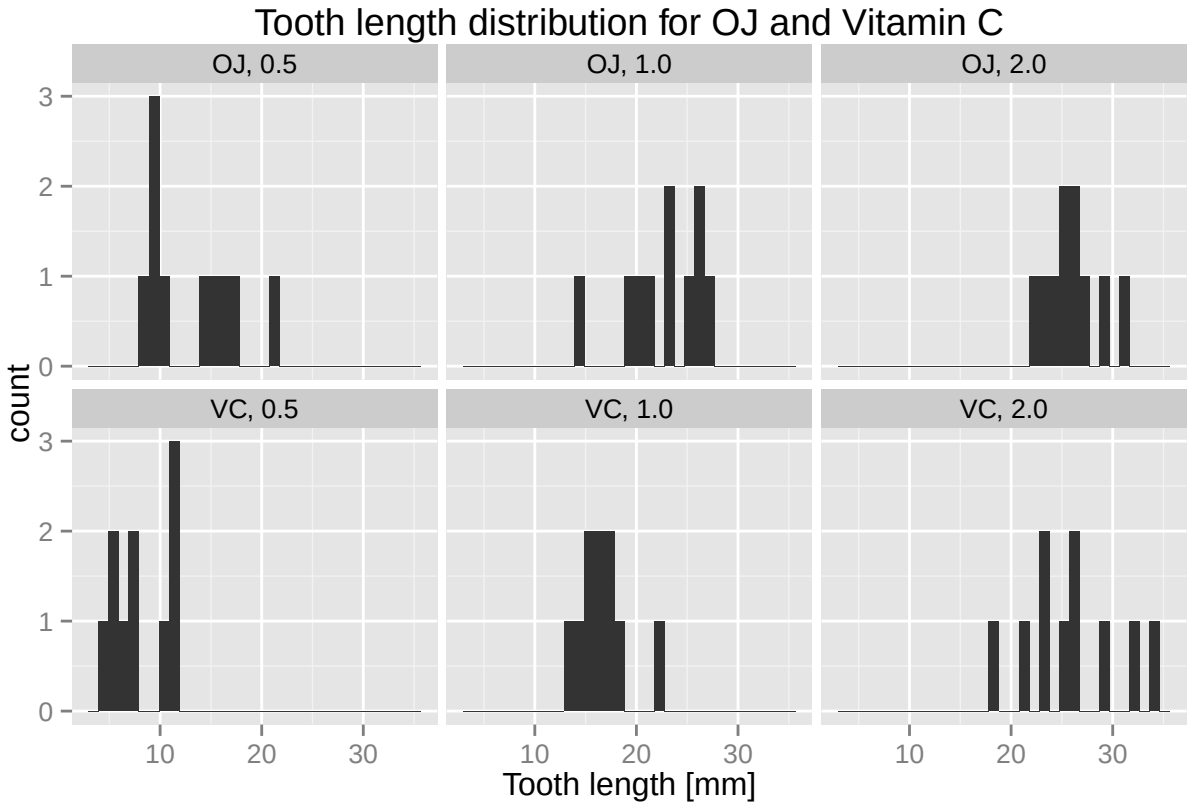
With this information we can draw a graph to illustrate how consumption of orange juice or vitamin C and the doses impact teeth growth. Let us first explore the data layout:

```r
library(ggplot2)
g <- ggplot(ToothGrowth, aes(x = dose, y = len))
g <- g + geom_point() +
     facet_wrap(~ supp) +
     labs(x = "Dose [mg]") +
     labs(y = "Tooth length [mm]") +
     labs(title = "Tooth length for OJ (Orange Juice) and Vitamin C")
g
```



Then we would check the data distribution by *supp* and *dose*. The distribution for orange juice is in the top row and for VC in the bottom row:

```r
g <- ggplot(ToothGrowth)
g <- g + geom_histogram(aes(x = len)) + facet_wrap(supp ~ dose) +
     labs(x = "Tooth length [mm]") +
     labs(title = "Tooth length distribution for OJ and Vitamin C")
g
```



Comparison of tooth growth by supp and dose

So we want to determine whether consumption of orange juice rather than vitamin C would increase teeth growth and whether it depends on the dose. We split *ToothGrowth* by doses and calculate the respective confidence intervals:

```
doses <- unique(ToothGrowth$dose)
```

The confidence interval for 0.5mg

```
d <- ToothGrowth[ ToothGrowth$dose == doses[ 1 ], ]
t.test(len ~ supp, data = d, paired = FALSE, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  len by supp
## t = 3.1697, df = 18, p-value = 0.005304
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.770262 8.729738
## sample estimates:
## mean in group OJ mean in group VC
##            13.23             7.98
```

The confidence interval for 1mg

```
d <- ToothGrowth[ ToothGrowth$dose == doses[ 2 ], ]
t.test(len ~ supp, data = d, paired = FALSE, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  len by supp
## t = 4.0328, df = 18, p-value = 0.0007807
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.840692 9.019308
## sample estimates:
## mean in group OJ mean in group VC
##            22.70            16.77
```

The confidence interval for 2mg

```
d <- ToothGrowth[ ToothGrowth$dose == doses[ 3 ], ]
t.test(len ~ supp, data = d, paired = FALSE, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  len by supp
## t = -0.0461, df = 18, p-value = 0.9637
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.722999  3.562999
## sample estimates:
## mean in group OJ mean in group VC
##            26.06            26.14
```

From the results printed I would say that for doses 0.5 and 1 orange juice seems to help growing longer teeth. We cannot say this for the 2 dose since the confidence interval spreads across zero rather symmetrically and the mean in the OJ group and in the VC group are almost the same.