Basic inferential data analysis

In this report we'll analyse ToothGrowth data set included in R and will try to do some inference

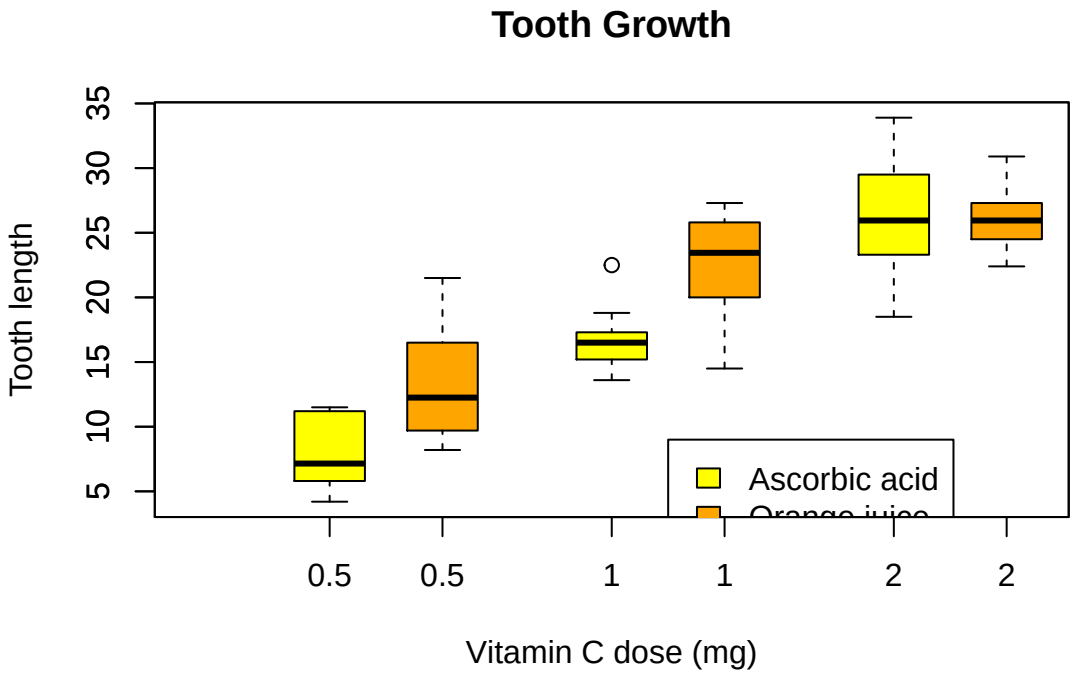
Loading data

```
data(ToothGrowth)
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 consists of 60 rows with 3 columns. First column (len) is measured value, second and third columns (supp and dose) are controlled variables.

Boxplota for *len* changing across *supp* and across *dose*:



Data summary

Let's calculate mean, median and standard deviation of data aggregated by *supp* and *dose*. Mean:

```
##   supp dose   len
## 1   OJ  0.5 13.23
## 2   VC  0.5  7.98
## 3   OJ  1.0 22.70
## 4   VC  1.0 16.77
## 5   OJ  2.0 26.06
## 6   VC  2.0 26.14
```

Median:

```
##   supp dose   len
## 1   OJ  0.5 12.25
## 2   VC  0.5  7.15
## 3   OJ  1.0 23.45
## 4   VC  1.0 16.50
## 5   OJ  2.0 25.95
## 6   VC  2.0 25.95
```

Standard deviation:

```
##   supp dose   len
## 1   OJ  0.5 4.460
## 2   VC  0.5 2.747
## 3   OJ  1.0 3.911
## 4   VC  1.0 2.515
## 5   OJ  2.0 2.655
## 6   VC  2.0 4.798
```

Inference

There are quite a few ways to subset data for mean comparisons. At first sight both *supp* and *dose* looks siginificant, so we'll compare acknowledging them. There are 9 possible comparisons in this case. To control for error rate we'll use Bonferroni correction. Confidence level will 0.95, giving 0.025 tail in both sides (upper and lower tail), using Bonferroni correction gives us confidence interval (0.0028, 0.9972).

Let's calculate p-value of each comparison.

Fixed *supp* = 'VC' supp = VC, dose 0.5 to 1

```
## [1] 6.811e-07
```

supp = VC, dose 0.5 to 2

```
## [1] 4.682e-08
```

supp = VC, dose 1 to 2

```
## [1] 9.156e-05
```

We can see that all three p-values are much lower than 0.0028, so *dose* really is significant when *supp* = "VC".

Fixed *supp* = 'OJ' supp = OJ, dose 0.5 to 1

```
## [1] 8.785e-05
```

supp = OJ, dose 0.5 to 2

```
## [1] 1.324e-06
```

supp = OJ, dose 1 to 2

```
## [1] 0.0392
```

First two p-values are again much lower than 0.0028, but third p-value is greater than 0.0028. So there is no statistically significant difference between *dose* equal 1 and 2 when *supp* is OJ.

Fixed *dose*, comparing *supp* dose = 0.5, supp VC to OJ

```
## [1] 0.006359
```

dose = 1, supp VC to OJ

```
## [1] 0.001038
```

dose = 2, supp VC to OJ

```
## [1] 0.9639
```

In this case p-value is lower than 0.0028 when *dose* equal 1, which says that there is a statistically significant difference between *supp* VC and OJ. In other two cases there is no statistically significant difference.

Summary

There are few assumptions we needed when performing inference:

- Population is normally distributed. That way we could use mean and standard deviation of sample dat as true population mean and standard deviation estimates
- Real population variance is the same for different *supp* and *dose*

Following are comparisons which proved to have statistically significant difference in average value:

- Fixed supp = VC, dose 0.5 to 1
- Fixed supp = VC, dose 0.5 to 2
- Fixed supp = VC, dose 1 to 2
- Fixed supp = OJ, dose 0.5 to 2
- Fixed supp = OJ, dose 1 to 2
- Fixed dose = 2, supp VC to OJ