Analysis of the ToothGrowth data

In this paper we will investigate ToothGrowth data (see dataset documentation). The data contains 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). A data frame with 60 observations on 3 variables:

- *len* (numeric) - tooth length,
- *supp* (factor) - supplement type (VC or OJ),
- *dose* (numeric) - dose in milligrams.

We will compare tooth growth with supplement type and dose by hypothesis tests.

Initial setup

Before we start data processing, we load necessary libraries for plots.

```
library(ggplot2)
library(GGally)
```

Loading and preprocessing the data

We perform a basic summary of the 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 ...
```

```
print(summary(ToothGrowth))
```

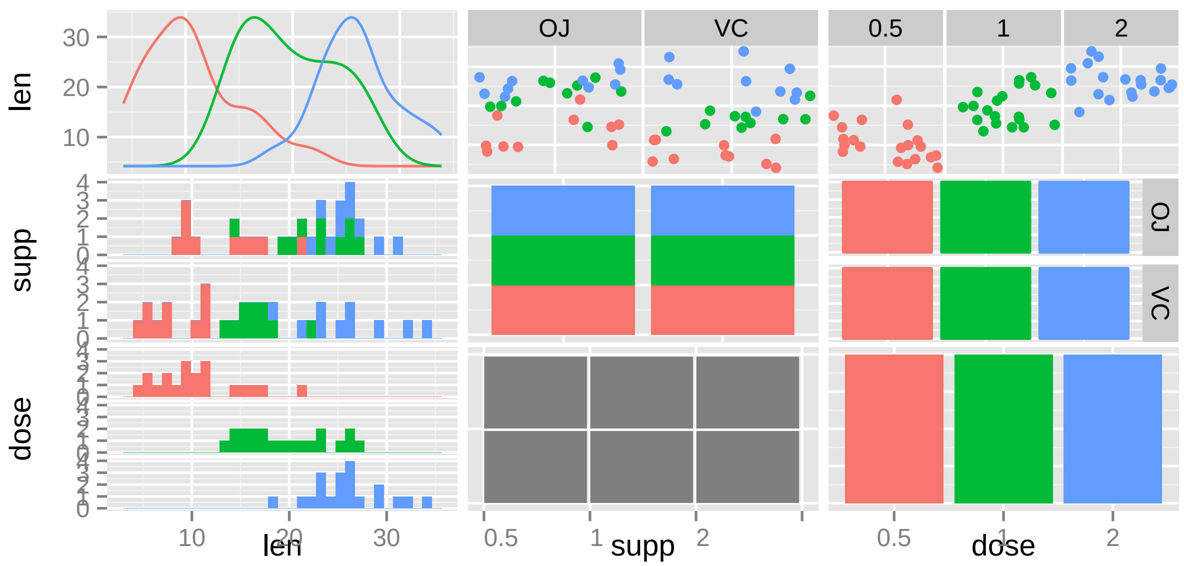
```
##       len        supp         dose      
##  Min.   : 4.20   OJ:30   Min.   :0.500  
##  1st Qu.:13.07   VC:30   1st Qu.:0.500  
##  Median :19.25           Median :1.000  
##  Mean   :18.81           Mean   :1.167  
##  3rd Qu.:25.27           3rd Qu.:2.000  
##  Max.   :33.90           Max.   :2.000  
```

As we can see, we have 3 attributes. The *dose* attribute is numeric but for analysis purpose it is better to turn it into a factor variable. The attribute has only 3 levels.

```
ToothGrowth$dose = factor(ToothGrowth$dose)
```

Let us perform some basic exploratory data analyses.

```
print(ggpairs(data=ToothGrowth,
              upper = list(continuous = "points",
                           combo      = "dot",
                           discrete   = "facetbar"),
              lower = list(continuous = "cor",
                           combo      = "facethist",
                           discrete   = "ratio"),
              colour="dose"))
```



As one can see, we have have the same proportions of data either we split it by *supp* or *dose* attribute. When we split the data by *supp*, it is hard to tell whether the means are the same. But when we split the data by *dose*, the average lenght of 3 groups seem to differ. We will focus on this problem in the next section.

Comparison of tooth growth by supp and dose

We use confidence intervals and hypothesis tests to compare tooth growth by supp and dose attributes. We perform paired t-tests on equality of means with 95% confidence level. Since the data contains results for 10 guinea pigs in each *supp* and *dose* combination, the t-test is paired. We can not say anything about equality of variances, so we assume that they are not equal. For each test a null hypothesis H_0 is that a difference in means is equal to 0; an alternative hypothesis H_a is that true difference in means is not equal to 0.

Firstly, we perform t-test when we split data by *supp* attribute.

```
TG_list = split(ToothGrowth, ToothGrowth$supp)
TGsOJ = TG_list[[1]]
TGsVC = TG_list[[2]]

t_s = t.test(x = TGsOJ$len, y = TGsVC$len,
             alternative = c("two.sided"),
             mu = 0, paired = TRUE, var.equal = FALSE,
             conf.level = 0.95)
```

Secondly, we perform t-tests when we split data by *dose* attribute.

```
TG_list = split(ToothGrowth, ToothGrowth$dose)
TGd05 = TG_list[[1]]
TGd10 = TG_list[[2]]
TGd20 = TG_list[[3]]

t_d0510 = t.test(x = TGd05$len, y = TGd10$len,
                 alternative = c("two.sided"),
                 mu = 0, paired = TRUE, var.equal = FALSE,
                 conf.level = 0.95)
t_d0520 = t.test(x = TGd05$len, y = TGd20$len,
                 alternative = c("two.sided"),
                 mu = 0, paired = TRUE, var.equal = FALSE,
                 conf.level = 0.95)
t_d1020 = t.test(x = TGd10$len, y = TGd20$len,
                 alternative = c("two.sided"),
                 mu = 0, paired = TRUE, var.equal = FALSE,
                 conf.level = 0.95)
```

Results and conclusions

Let us look on the result of the t-test in by-*supp* split groups:

```
print(c(t_s$conf.int[1:2], t_s$p.value))
```

```
## [1] 1.408658641 5.991341359 0.002549842
```

The interval is entirely above zero. The p-value is lower than 0.05. There is sufficient evidence to reject H_0, so we reject H_0 in favor of H_a.

Now let us look on the results of the t-tests in by-*dose* split groups:

```
print(c(t_d0510$conf.int[1:2], t_d0510$p.value))
```

```
## [1] -1.187288e+01 -6.387121e+00  1.225437e-06
```

```
print(c(t_d0520$conf.int[1:2], t_d0520$p.value))
```

```
## [1] -1.836720e+01 -1.262280e+01  7.190255e-10
```

```
print(c(t_d1020$conf.int[1:2], t_d1020$p.value))
```

```
## [1] -9.2581856558 -3.4718143442  0.0001934186
```

All intervals are entirely below zero. In each case p-value is lower than 0.05. In each case there is sufficient evidence to reject H_0, so in each case we reject H_0 in favor of H_a.

To conclude, the interpreation of results is that the length of teeth depends on dose level of Vitamin C and on a delivery method.