Statistical Inference - Course project 2 (Author: Peter)

Part II

In the second part I will analyze the ToothGrowth data set that comes with the R `datasets` package, performing some basic exploratory analysis, data summary and using confidence intervals for further statistical analysis.

Preliminaries

Let us first of all load packages and data.

```
library(ggplot2)
library(gridExtra)
library(datasets)
# now load data
data("ToothGrowth")
tooths = ToothGrowth
tooths$dose = as.factor(tooths$dose)
```

Basic summary and exploratory analysis of the data

As often, we simply start out with printing out some information about the data, the ones given by `str`, `summary`, and a simple boxplot.

```
str(tooths)
```

```
## '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: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

```
summary(tooths)
```

```
##       len        supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
##  Median :19.25           2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

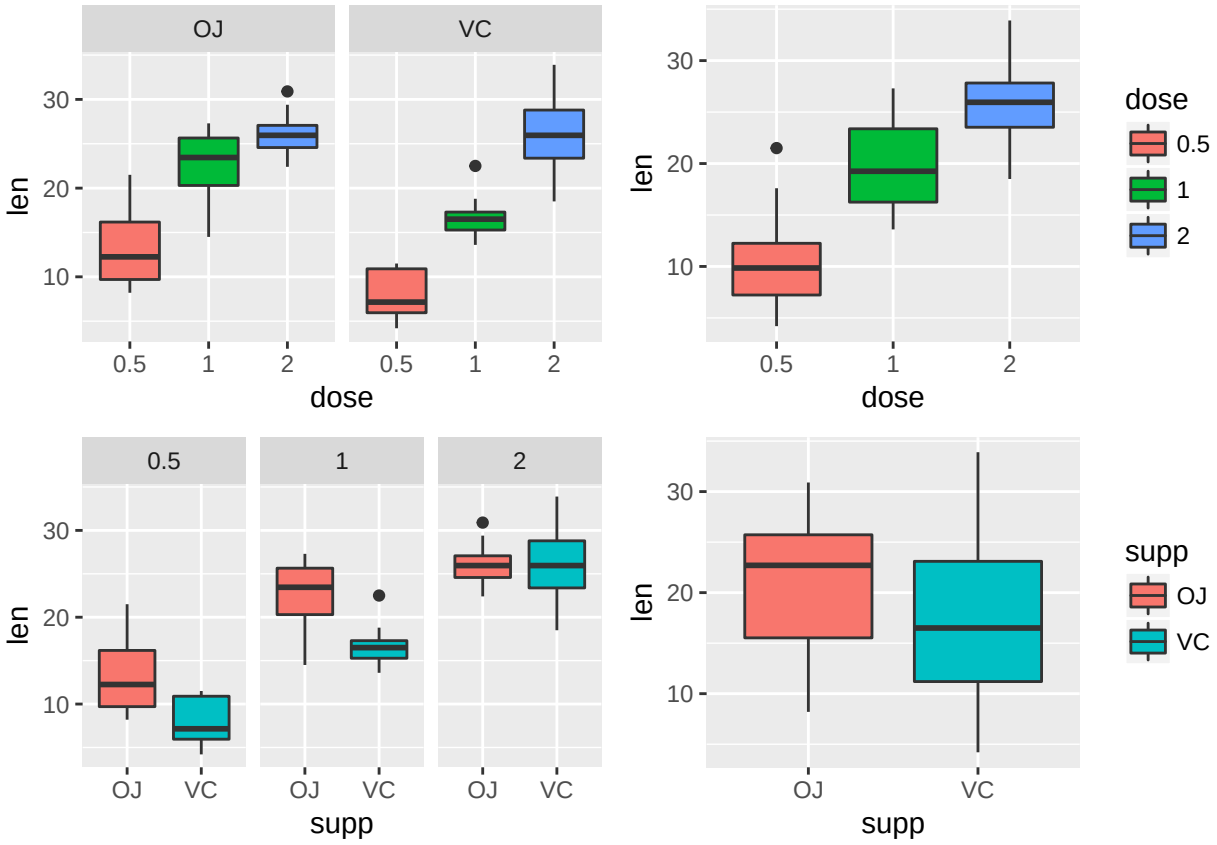
```
p1 <- ggplot(data=tooths, aes(x=dose,y=len,fill=dose)) +
  geom_boxplot() +
  theme(legend.position="none") +
  facet_grid(.~supp)

p2 <- ggplot(data=tooths, aes(x=supp,y=len,fill=supp)) +
  geom_boxplot() +
  theme(legend.position="none") +
  facet_grid(.~dose)

p3 <- ggplot(data=tooths, aes(x=supp,y=len,fill=supp)) +
  geom_boxplot()
```

```
p4 <- ggplot(data=tooths, aes(x=dose,y=len,fill=dose)) +
  geom_boxplot()

grid.arrange(p1, p4, p2, p3, ncol = 2, nrow=2)
```



Hypothesis tests with confidence intervals

Let's start with the "null" hypothesis that there is no correlation between the `delivery method` and the `tooth length`, and let's analyze possible correlation between the two.

```
t.test(len ~ supp, paired=FALSE, var.equal=FALSE, data=tooths)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by 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
```

The test tells us the 95% confidence interval is $[-0.1710156, 7.5710156]$, which contains the zero value. Furthermore, the p-value is 0.06 that is larger than 0.05, therefore we cannot reject the null hypothesis.

Let's double check this result with the analysis of variance (ANOVA):

```
anovaRes = aov(len ~ supp*dose, data=tooths)
summary(anovaRes)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## supp         1  205.4   205.4  15.572 0.000231 ***
## dose         2 2426.4  1213.2  92.000  < 2e-16 ***
## supp:dose    2  108.3    54.2   4.107 0.021860 *
## Residuals   54  712.1    13.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This result show that there is a clear correlation between the `len` (length) and the `dose` (dosage) parameters, as well has between `len` and `supp` (Supplement), and the two conclusions are independent.

Conclusions

The analysis perfomed has shown that

1. A correlation between the Delivery Method and the Tooth Length is possible (we could not rule out the null hypothesis)
2. Both the Supplement and the Dosage have independent effect on the Tooth Length.

It is worth noticing that these results are valid under the following assumptions:

- The sampled data are representative of the population (of the guinea pigs)
- Dosage and Supplement were randomly distributed in the population
- The means' distribution is normal.