Inferential data analysis

Akim van Eersel

2020-10-11

Exploratory data analysis : tooth growth data

The data set 'ToothGrowth' is aiming to study the effect of vitamin C on tooth growth in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice or ascorbic acid (a form of vitamin C and coded as VC).

Load the data

```
library(datasets)

# Load the data
data(ToothGrowth)

# Store the data set in a variable
tooth <- ToothGrowth

# Rename some values to explicit labeling
tooth$supp <- factor(tooth$supp, labels= c("orange juice", "ascorbic acid"))
colnames(tooth) <- c('len', 'delivery.method', 'dose')
```

Statistics summary

Get the dimensions of the data set and by features.

```
# Number of rows and columns
dim(tooth)
```

```
## [1] 60  3
```

```
# Number of observations for the "dose" and "delivery.method" variables
table(tooth$dose, tooth$delivery.method)
```

```
##
##       orange juice ascorbic acid
##   0.5           10            10
##   1             10            10
##   2             10            10
```

Let's see the 5 first rows of the data set.

```
head(tooth)
```

```
##    len delivery.method dose
## 1  4.2   ascorbic acid  0.5
## 2 11.5   ascorbic acid  0.5
## 3  7.3   ascorbic acid  0.5
## 4  5.8   ascorbic acid  0.5
## 5  6.4   ascorbic acid  0.5
## 6 10.0   ascorbic acid  0.5
```

Overall statistics on the length of the tooth growth.

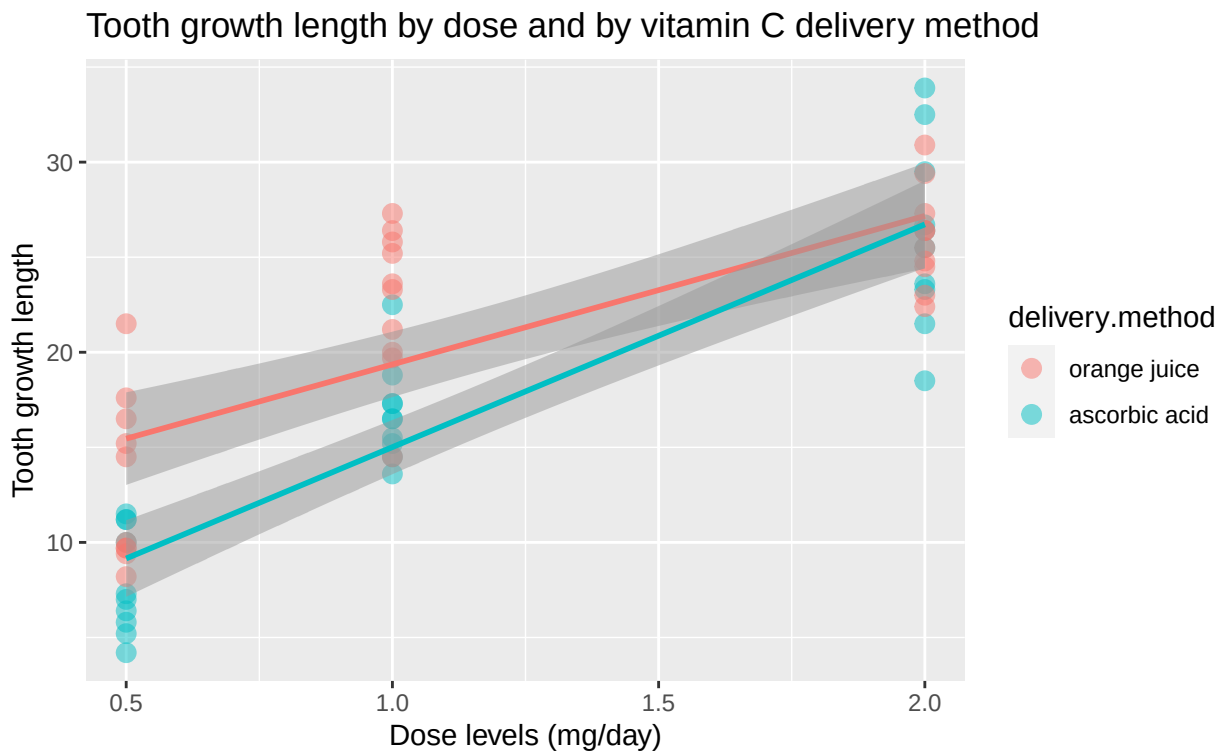
```
# Tukey's five number summary + mean & std
data.frame(Statistic = c('Min', '25th percentile','Median', '75th percentile', 'Max', 'Mean', 'Standard
  Values = c(fivenum(tooth$len), mean(tooth$len), sd(tooth$len)))
```

```
##            Statistic    Values
## 1                Min  4.200000
## 2    25th percentile 12.550000
## 3             Median 19.250000
## 4    75th percentile 25.350000
## 5                Max 33.900000
## 6               Mean 18.813333
## 7 Standard Deviation  7.649315
```

Now, let's see the length of tooth growth discriminated by other variable in a plot.

```
library(ggplot2)

ggplot(data = tooth, aes(dose, len, color = delivery.method), ) +
      geom_point(size = 3, alpha = 1/2) +
      geom_smooth(method = "lm", alpha =1/2, show.legend = F)  +
      labs(title = "Tooth growth length by dose and by vitamin C delivery method",
           x = "Dose levels (mg/day)", y = "Tooth growth length")
```



The plot is showing all data points and a linear regression of the length function of the dose levels for each vitamin C delivery method.

Inferential analysis

Hypothesis testing

Two hypothesis are set :
1. Whatever the delivery method, tooth growth length is increasing as the increase of the dose levels.
2. For both 0.5 and 1.0 mg/day dose levels delivery method is more effective with the orange juice.

Length increase with dose levels

In order to reduce the amount of calculations, we'll only focus on dose levels 0.5 and 2.0 mg/day. In that sense :
* Null hypothesis, H0 : means of the length variable are equal, mu_2.0 - mu_0.5 = 0 * Alternative hypothesis, Ha : means of the length variable are different, mu_2.0 - mu_0.5 > 0

we assume the normality of the distribution of length variable on both dose levels, also as the independence between subjects. Similar variation is observed between the two groups.
A t-test will be conducted.

```
g2 <- tooth$len[tooth$dose == 2]
g1 <- tooth$len[tooth$dose == 0.5]

t.test(g2-g1, var.equal = TRUE, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  g2 - g1
## t = 11.291, df = 19, p-value = 3.595e-10
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  13.12216      Inf
## sample estimates:
## mean of x
##    15.495
```

Even if t-test's results explicitly state that the null hypothesis is rejected, the related p-value less than 0.01 and the 95% confidence interval of the means difference not including 0 confirm H0 rejection.

In conclusion, whatever the delivery method, tooth growth length is higher with a 2.0 mg/day dose level than a 0.5 mg/day dose level.

Greater length with on delivery method

In order to reduce the amount of calculations, we'll only focus on dose level 1.5 mg/day. Data on the 1.0 mg/day dose level by delivery method overlap on each other but still seems to be enough differentiated, let's investiguate. In that sense :
* Null hypothesis, H0 : means of the length variable are equal, mu_orange.juice - mu_ascorbic.acid = 0 * Alternative hypothesis, Ha : means of the length variable are different, mu_orange.juice - mu_ascorbic.acid > 0

we assume the normality of the distribution of length variable on both dose levels, also as the independence between subjects. Similar variation is observed between the two groups.
A t-test will be conducted.

```
g2 <- tooth$len[tooth$dose == 1 & tooth$delivery.method == "orange juice"]
g1 <- tooth$len[tooth$dose == 1 & tooth$delivery.method == "ascorbic acid"]

t.test(g2-g1, var.equal = TRUE, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  g2 - g1
## t = 3.3721, df = 9, p-value = 0.004115
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  2.706401      Inf
## sample estimates:
## mean of x
##      5.93
```

Again, even if t-test's results explicitly state that the null hypothesis is rejected, the related p-value less than 0.01 and the 95% confidence interval of the means difference not including 0 confirm H0 rejection.

In conclusion, for a 1.0 mg/day dose level, tooth growth length is higher with orange juice vitamin C delivery method than ascorbic acid vitamin C delivery method.