Statistical Inference Course Project Part 2

rcblt@github.com

Overview

In this report we analyzed the ToothGrowth dataset to test whether different doses and supplement types had an influence on the odontoblast lengths. Using a paired T-Test we showed that -as initially seen from an exploratory data analysis- statistically significant differences in lengths are present for different doses and supplements. In addition we computed the measurable minimum difference in length for a given statistical power of 80%.

Loading and cleaning data

```
library(datasets)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Data summary and explorative data analysis

The `TootGrowth` dataset is a dataframe containing 60 observations (rows) of 3 variables (columns).

```
head(ToothGrowth)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```
dim(ToothGrowth)
```

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

Variables `len` and `dose` are numeric while `supp` is encoded as factor. From the ToothGrowth dataset documentation we see that the values `VC` and `OJ` represent the delivery method of a supplement and stand for vitamin C and orance juice respectively. Also, values of dose are expressed in milligrams/day.

```
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 ...
```

Summary of the `TootGrowth` dataset. There are 30 observations for each supplement.

```
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
```

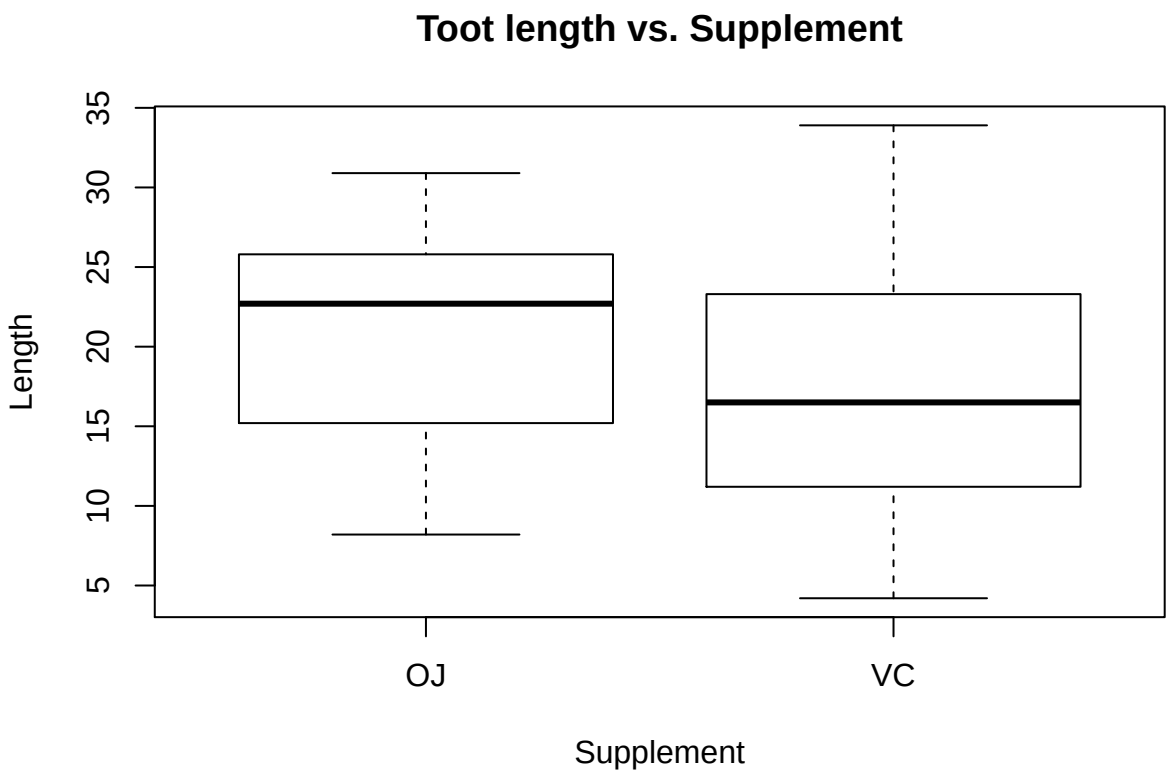
Mean values and standard deviation of the length variable grouped by supplement and dose

```
ToothGrowth %>%
    group_by(supp, dose) %>%
    summarise(Mean= mean(len), SD= sd(len), N= n())
```

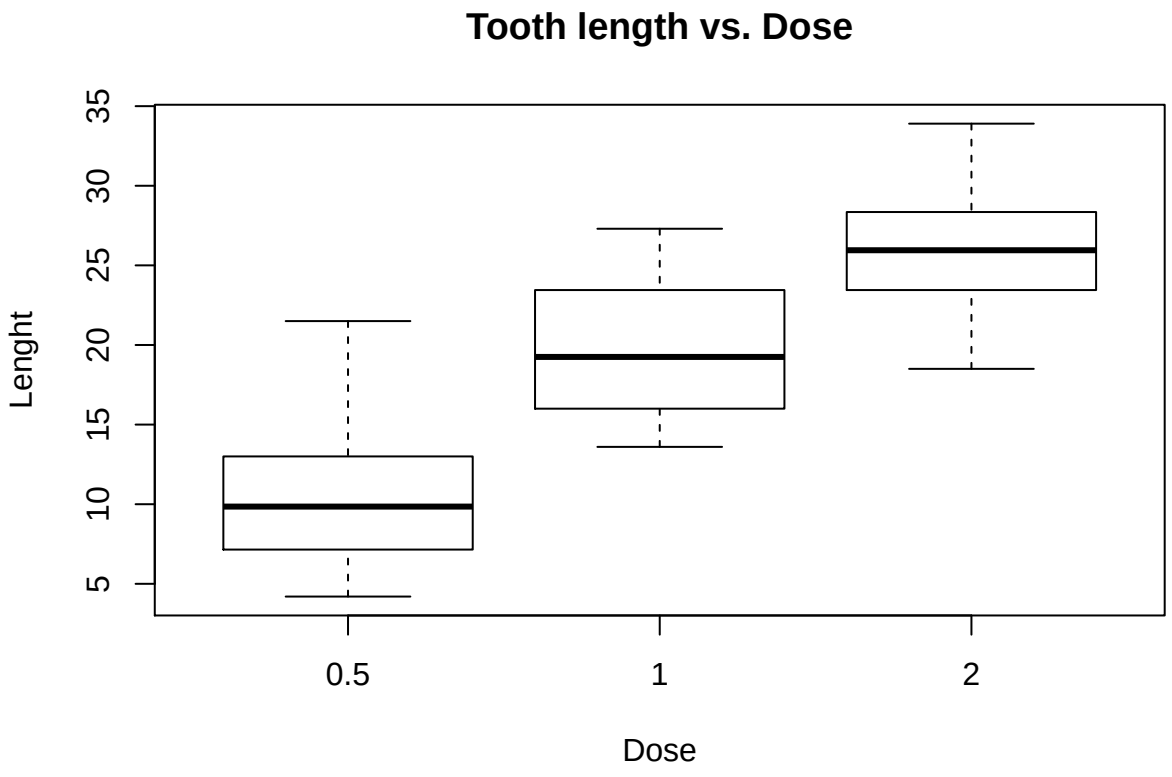
```
## Source: local data frame [6 x 5]
## Groups: supp
##
##   supp dose  Mean       SD  N
## 1   OJ  0.5 13.23 4.459709 10
## 2   OJ  1.0 22.70 3.910953 10
## 3   OJ  2.0 26.06 2.655058 10
## 4   VC  0.5  7.98 2.746634 10
## 5   VC  1.0 16.77 2.515309 10
## 6   VC  2.0 26.14 4.797731 10
```

Odontoblasts (cells responsible for tooth growth) lengths versus supplement and dose

```
boxplot(ToothGrowth$len ~ ToothGrowth$supp,
        xlab= "Supplement",
        ylab= "Length",
        main= "Toot length vs. Supplement")
```




```
boxplot(ToothGrowth$len ~ ToothGrowth$dose,
        xlab= "Dose",
        ylab= "Lenght",
        main= "Tooth length vs. Dose")
```



From the exploratory analysis above we can see a potential effect of dose on the odontoblast lenght while the effect of supplement looks less clear. However, looking at the mean values of odontoblast lengths grouped by supplement and dose one can see that also supplement may have an effect for a fixed dose value.

Hypothesis testing of odontoblast growth by supplement and dose

Assuming a normal distribution of odontoblast lengths and common variances among groups we want to test whether each treatement alone has an effect on average length. So, for example, we want to see if the diffence of the mean lengths at doses 0.5 and 1 when the orange juice supplement was used is significatively different from zero.

So, our assumptions are:

1. sample lengths are iid following normal distribution
2. common variance between groups of samples
3. a Type I error of 0.05

As we have 10 samples for each group we will use a two-sided T-test to evaluate whether the statistics:

$$\frac{\mu_B - \mu_A}{SE}$$

is different from zero, where μ_A and μ_B are the mean lengths of the two groups under testing.

In our example above A is the group of samples at 0.5 mg/day and B at 1.0 mg/day. Both with the OJ supplement.

Retrieve the values for the two conditions under test from the dataset and execute the test.

```
A <- as.vector(ToothGrowth %>% filter(supp== "OJ", dose== 0.5) %>% select(len))
B <- as.vector(ToothGrowth %>% filter(supp== "OJ", dose== 1.0) %>% select(len))
t.test(B, A, var.equal = TRUE, conf.level = 0.95)
```

```
##
##  Two Sample t-test
##
## data:  B and A
## t = 5.0486, df = 18, p-value = 8.358e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   5.529186 13.410814
## sample estimates:
## mean of x mean of y
##     22.70     13.23
```

Let's see if supplement has an effect at a given dose.

```
A <- as.vector(ToothGrowth %>% filter(supp== "OJ", dose== 0.5) %>% select(len))
B <- as.vector(ToothGrowth %>% filter(supp== "VC", dose== 0.5) %>% select(len))
t.test(B, A, var.equal = TRUE, conf.level = 0.95)
```

```
##
##  Two Sample t-test
##
## data:  B and A
## t = -3.1697, df = 18, p-value = 0.005304
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.729738 -1.770262
## sample estimates:
## mean of x mean of y
##      7.98     13.23
```

Given the sample groups above let's compute which is the minimum difference in length we can detect with a probability of at least 80% (power= 0.8):

```
sd <- sqrt(((nrow(A)-1)*(sd(A$len)^2)+(nrow(B)-1)*(sd(B$len)^2))/(nrow(A)+nrow(B)-2))
power.t.test(n= nrow(A), sd = sd, sig.level = 0.05, power = 0.8, type = "two", alternative = "two")
```

```
##
##      Two-sample t test power calculation
##
##               n = 10
##           delta = 4.907041
##              sd = 3.703579
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

See the Results section for the full panel of tests.

Discussion and conclusions

Formal hypothesis testing using a T-test for indipendent groups of samples under the assumption that distribution is normal and groups have the same variance confirmed that dose has always direct impact on odontoblast length. This holds true with very high statistical significance in all cases except one. In the case of orange juice supplement comparison of 2.0 mg/day vs. 1.0 mg/day we obtained a PValue of 0.037 that is not very far from the limit of the acceptable error rate, hence there is no real strong evidence that 2.0 mg/day had a an impact on length with respect to 1.0 mg/day. Comparison of the two supplements at the same dosage gave generally higher PValues and except in the case of 2.0 mg/day where we can't see any significant differece among the two types of supplement. In fact, mean values are almost identical in this case. We may suppose that 2 mg/day is close to the level of "saturation" of the supplement so differences a further dosage increase and in type of supplement became not relevant in terms of odontoblast length.

We also investigated the power of our tests, in particular we computed the minimum difference in odontoblast length we can detect with a power of 80% ($1 - \beta = 0.8$). Given the relatively low number of samples per group (10) we obtained a quite high values, sometimes as high as half the mean length of the group. In other words we may state that given the low number of samples we do not have much power to observe small differences in odontoblast lengths with a high (80%) level of confidence.

In conclusion, we showed that under the given assumption we can tell that both supplement and dose have a statistically significant effect on the odontoblast lenghts according to the sample data we analyzed.

Results

We repeated the 95% confidence level T-tests for all combinations of doses and supplements.

The following tables summarises the results obtained.

- `Param`: the fixed parameter used for this test among all possible values of supplements and doses
- `Test`: the specific comparison under test, for examples 0.5 mg/day versus 1 mg/day
- `PValue`: the PValue of the test statistics
- `MeanA`: the mean value of condition A (i.e. null hypothesis)
- `MeanB`: the mean value of condition B (i.e. alternative hypothesis)
- `CILow`: the low limit of the 95% confidence interval
- `CIHigh`: the high limit of the 95% confidence interval
- `T`: the value of the test statistics
- `MinDist80`: the minimum difference in odontoblast length that can be notices with a power of 80%

```
##   Param     Test   PValue MeanA MeanB   CILow  CIHigh       T MinDist80
## 1    OJ 0.5vs1.0 0.000084 13.23 22.70  5.5292 13.4108  5.0486    5.5572
## 2    OJ 0.5vs2.0 0.000000 13.23 26.06  9.3818 16.2782  7.8170    4.8626
## 3    OJ 1.0vs2.0 0.037363 22.70 26.06  0.2195  6.5005  2.2478    4.4287
## 4    VC 0.5vs1.0 0.000001  7.98 16.77  6.3157 11.2643  7.4634    3.4893
## 5    VC 0.5vs2.0 0.000000  7.98 26.14 14.4872 21.8328 10.3878    5.1794
## 6    VC 1.0vs2.0 0.000034 16.77 26.14  5.7710 12.9690  5.4698    5.0752
## 7   0.5   OJvsVC 0.005304 13.23  7.98 -8.7297 -1.7703 -3.1697    4.9070
## 8   1.0   OJvsVC 0.000781 22.70 16.77 -9.0193 -2.8407 -4.0328    4.3565
## 9   2.0   OJvsVC 0.963710 26.06 26.14 -3.5630  3.7230  0.0461    5.1373
```