ToothGrowth Data Analysis

ThinkersPark

2023-02-28

ToothGrowth data load & summary

This report presents Part 2 of the Statistical Inference project: Analysis of the ToothGrowth data.

First, data is loaded and basic features summarised:

```
library(datasets)
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
```

Type help("ToothGrowth") for details, description quoted here: *The response is the length of odontoblasts (cells responsible for 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).* The below plot provides a visual illustration.

It is clear that increasing the dose of each supplement increases the tooth growth. So one could ask the question **"Which supplement is more effective?"/ "Does any of the supplements increase the tooth growth faster?"**

```
library(ggplot2)
g <- ggplot(ToothGrowth, aes(dose, len))
g <- g + geom_point() + geom_smooth(method="lm") + facet_grid(.~supp)
g <- g + labs(title="Length of odontoblasts by dose of supplement,
for OJ (orange juice) and VC (vitamin C) supplements.",
   x="Dose [mg/day]",
   y="Length of odontoblasts [mm]")
g
```

```
## `geom_smooth()` using formula 'y ~ x'
```

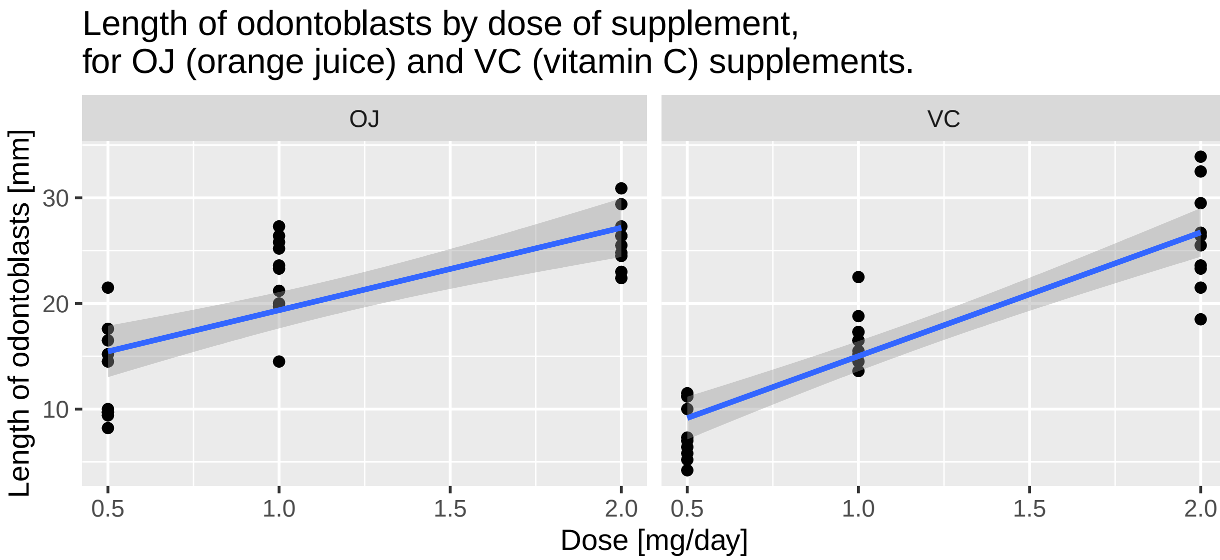

Figure 1: Length of odontoblasts by dose of supplement, for OJ (orange juice) and VC (vitamin C) supplements

Tooth growth by supplement and dose - comparison

In this case, it makes sense to analyse tooth growth (measured by length of ondoblasts in mm) by dose (measured in mg/ day) - denoted by a new variable "lenperdose" - and compare it between groups of animals given different supplements. For each group, the sample size (30 each), sample mean and standard deviation is calculated using the below code:

```
tg <- transform(ToothGrowth, lenperdose = len/dose)
n_oj <- length(tg$lenperdose[which(tg$supp=="OJ")])
n_vc <- length(tg$lenperdose[which(tg$supp=="VC")])
m_oj <- mean(tg$lenperdose[which(tg$supp=="OJ")])
m_vc <- mean(tg$lenperdose[which(tg$supp=="VC")])
sd_oj <- sd(tg$lenperdose[which(tg$supp=="OJ")])
sd_vc <- sd(tg$lenperdose[which(tg$supp=="VC")])
mdiff <- m_oj-m_vc; sddiff <- sqrt((sd_oj)^2/n_oj+(sd_vc)^2/n_vc)
df <- ((sd_oj)^2/n_oj+(sd_vc)^2/n_vc)^2/((sd_oj^2/n_oj)^2/(n_oj-1)+(sd_vc^2/n_vc)^2/(n_vc-1))
print(paste("mdiff=",round(mdiff,4),", sddiff=",round(sddiff,4),", df=",round(df,4)))
```

```
## [1] "mdiff= 5.4633 , sddiff= 1.6208 , df= 42.6098"
```

Assumptions

Per data description, the supplement groups are independent, comprising of 30 animals each, given different supplement doses (not time-dependent/ not paired); measurements are assumed i.i.d. (independent/ identically distributed); the distribution of each measurement is in fact unknown (rather not normal), and so the t-distribution used below is an approximation.

The sample size (for each group) is rather small, we therefore assume the difference between the two means to follow (approximately) a t-distribution:

$m_{diff} \sim t(0, s_{diff}, df)$, where:

- $m_{diff} = m_{oj} - m_{vc}$ denotes the difference between the average tooth growth per dose in each supplement group,
- $s_{diff} = \left(\frac{s_{oj}^2}{n_{oj}} + \frac{s_{vc}^2}{n_{vc}}\right)^{1/2}$, denotes the standard deviation of m_{diff}, and
  s_{oj} and s_{vc} are standard deviations of tooth growth per dose, in "orange juice" and "vitamin c" group, respectively,
- n_{oj} and n_{vc} are sizes of each group, both equal to 30,
- $df = \frac{\left(s_{oj}^2/n_{oj} + s_{vc}^2/n_{vc}\right)^2}{\left(\frac{s_{oj}^2}{n_{oj}}\right)^2/(n_{oj}-1) + \left(\frac{s_{vc}^2}{n_{vc}}\right)^2/(n_{vc}-1)}$.

Confidence intervals & hypothesis testing: Which supplement is more effective?

In this section, we will test the null hypothesis that the difference between the average tooth growth per dose in each supplement group is zero,

$H_0 : m_{diff} = 0$, equivalent to $H_0 : m_{oj} - m_{vc} = 0$,

vs. (one sided) alternative hypothesis that in fact the average tooth growth per dose in the "orange juice" group is higher:

$H_1 : m_{diff} > 0$, equivalent to $H_1 : m_{oj} > m_{vc}$.

First, let us note that the (two-sided) 95% confidence interval for m_{diff} is +/-3.2696; with m_{diff} value equal to 5.4633, it is an indication that the difference between the average tooth growth per dose in each supplement group is in fact non-zero, with high confidence.

```
l_2.5 <- qt(0.025,df)*sddiff; u_2.5 <- qt(0.975,df)*sddiff;
print(paste("lower_2.5% quantile=",round(l_2.5,4)," upper_2.5% quantile=",round(u_2.5,4)))
```

```
## [1] "lower_2.5% quantile= -3.2696  upper_2.5% quantile= 3.2696"
```

For actual testing of the hypothesis, we will use R function *t.test()*, noting that:

$m_{diff} \sim t(0, s_{diff}, df)$ is equivalent to $m_{diff}/s_{diff} \sim t(0, 1, df)$, which is the statistic displayed in the t.test() results.

```
pval <- t.test(tg$lenperdose[which(tg$supp=="OJ")],tg$lenperdose[which(tg$supp=="VC")],
       alternative="greater", paired=FALSE,var.equal=FALSE)$p.value
pow <- power.t.test(n=30,delta=mdiff,sd=sddiff,type ="two.sample",alt ="one.sided")$power
print(paste("T-test p-value=",round(pval,4)," T-Test power=",pow))
```

```
## [1] "T-test p-value= 8e-04  T-Test power= 1"
```

Conclusion

With p-value below 0.1% (probability of type 1 error/ rejecting null when it is in fact true), the test shows that **with a very high confidence level - above 99.9% - the null hypothesis can be rejected, and the average tooth growth per dose in the "orange juice" group is higher**.

Additionally, the power of the test is calculated using R function *power.t.test()*. With power (almost) equal to 1 (almost 0% probability of type 2 error/ failing to reject null when it is in fact false), the test demonstrates significant power, confirming the previous observation.