Tooth Growth Analysis

Americo

19 luglio 2015

## Packages

require(dplyr)  
require(ggplot2)  
require(datasets)  
require(tidyr)  
require(grid)  
require(gridExtra)

## Goals

We're going to analyze the ToothGrowth data in the R datasets package to perform some basic exploratory data analyses. Goals are to: \* provide a basic summary of the data; \* use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose (we'll only use the techniques from class, even if there's other approaches worth considering); \* state conclusions and the assumptions needed for conclusions.

## Dataset

Loading the dataset and studying the variables through the codebook and some initial exploration.

data(ToothGrowth)  
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

tail(ToothGrowth)

## len supp dose  
## 55 24.8 OJ 2  
## 56 30.9 OJ 2  
## 57 26.4 OJ 2  
## 58 27.3 OJ 2  
## 59 29.4 OJ 2  
## 60 23.0 OJ 2

dim(ToothGrowth)

## [1] 60 3

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 ...

myToothGrowth <- ToothGrowth

The dataset is composed by measurements on 10 pigs regarding the length of teeth (variable len) after the somministration of three different dose of vitamin (the variable dose) and two delivery methods (orange juice or ascorbic acid - the variable supp).

## Exploratory data analysis

How delivery methods and doses are distributed?

There is not a column dedicated to the ID of the ten pigs, so I assume that the measurements are ordered this way: each ten observations represents a dose of vitamin, the first 30 with a delivery methods and the second 30 with the other one.

ggplot(data = myToothGrowth, aes(x = supp, fill = factor(dose))) +  
 geom\_bar(width = 0.5) +  
 guides(fill=guide\_legend(reverse=TRUE)) +  
 ggtitle("distribution of doses and delivery methods")



My assumption (*I assume that the measurements are ordered this way: each ten observations represents a dose of vitamin, the first 30 with a delivery methods and the second 30 with the other one*) seem confirmed by the plot.

Let'see the distribution of teeth's lengths.

ggplot(data = myToothGrowth, aes(x = len)) +  
 geom\_histogram(binwidth = 1, colour = "black", fill = "gray") +  
 ggtitle("Histogram of teeth length")

I don't know if this kind of exploration makes sense, because this 60 observations are not really *a population*, are much more the sum of 6 different "snapshot"" of a population. Let's plot them so.

ggplot(data = myToothGrowth, aes(x = factor(dose), y = len)) +  
 geom\_boxplot() +  
 facet\_wrap(~supp) +  
 ggtitle("Boxplot of teeth length")

It seems that dose influences the growth of lenghts (very expected!) and that ascorbic acide is more effective than orange juice in facilitate this. Almost all population are quiete symmetrical, which is good for our further analysis. **Infact we are deailing with paired small samples**, so it's likely we are going to use t-test to compare groups, and t-test needs quite normal population to be accurate. Of course we can't check the normality of population (unless we use some tools that were not taugth in the class, but students have been explicitly discouraged in doing so), so we must perform analysis on these small samples. Furthermore, being paired samples, it's better to plot the differences between lengths, as we will test them with the t distribution. I will do this in the statistical analysis chapter, after some data manipulation.

## Data manipulation

I want to create a dataframe with 7 variables, the 6 measurements and the ID of the subjects. It seems to me the better way to flexibly perform t-tests.

mydfs <- list()  
j <- 1:10  
for (i in 1:6) {  
 mydfs[[i]] <- myToothGrowth[j,]  
 names(mydfs[[i]]) <- c(paste("len", i), paste("supp", i),paste("dose", i))  
 j <- j + 10  
}  
myToothGrowth2 <- bind\_cols(mydfs)  
myToothGrowth2$pig\_id <- 1:10  
  
#renaming variables to have a smaller dataframe  
myToothGrowth3 <- myToothGrowth2[, seq(1, 19, 3)]  
names(myToothGrowth3) <- c("lenVC\_05", "lenVC\_1", "lenVC\_2", "lenOJ\_05", "lenOJ\_1", "lenOJ\_2", "pig\_id")  
head(myToothGrowth3)

## lenVC\_05 lenVC\_1 lenVC\_2 lenOJ\_05 lenOJ\_1 lenOJ\_2 pig\_id  
## 1 4.2 16.5 23.6 15.2 19.7 25.5 1  
## 2 11.5 16.5 18.5 21.5 23.3 26.4 2  
## 3 7.3 15.2 33.9 17.6 23.6 22.4 3  
## 4 5.8 17.3 25.5 9.7 26.4 24.5 4  
## 5 6.4 22.5 26.4 14.5 20.0 24.8 5  
## 6 10.0 17.3 32.5 10.0 25.2 30.9 6

## Statistical analysis

As seen before, we are dealing witg 6 measurements on 10 subjects. These measurements were taken in 2 different conditions, let'say: *Orange juice* and *Ascorbic acid*. So we must be very careful in comparing two measurements belonging to different conditions.

My idea is:

* to compare lengths derived from different doses given with the same delivery method, to see if bigger dose (as we expect) related to bigger length;
* to compare same doses given with different delivery methods, to see if different delivery method relates to different length (looking at the boxplots, i think it worth considering the dose 0.5 and 1, which seems the ones with bigger difference between lengths distributions and their averages).

But first of all I need to perform some check on normality of differences, although the size is very very small. Naming convention is: VC\_1\_05 = supp VC dose 1 - supp Vc dose 0.5 and VC\_OJ\_05 = supp OJ - supp VC for dose 2.

### Verification of normality assumptions

First of all I will create a dataframe of the differences between all the vectors I aim to compare.

myToothGrowth\_diff <- data.frame (  
 lenVC\_1\_05 = myToothGrowth3$lenVC\_1 - myToothGrowth3$lenVC\_05,  
 lenVC\_2\_1 = myToothGrowth3$lenVC\_2 - myToothGrowth3$lenVC\_1,  
 lenVC\_2\_05 = myToothGrowth3$lenVC\_2 - myToothGrowth3$lenVC\_05,  
 lenOJ\_1\_05 = myToothGrowth3$lenOJ\_1 - myToothGrowth3$lenOJ\_05,  
 lenOJ\_2\_1 = myToothGrowth3$lenOJ\_2 - myToothGrowth3$lenOJ\_1,  
 lenOJ\_2\_05 = myToothGrowth3$lenOJ\_2 - myToothGrowth3$lenOJ\_05,  
 lenVC\_OJ\_05 = myToothGrowth3$lenOJ\_05 - myToothGrowth3$lenVC\_05,  
 lenVC\_OJ\_1 = myToothGrowth3$lenOJ\_1 - myToothGrowth3$lenVC\_1,  
 lenVC\_OJ\_2 = myToothGrowth3$lenOJ\_2 - myToothGrowth3$lenVC\_2)  
  
head(myToothGrowth\_diff)

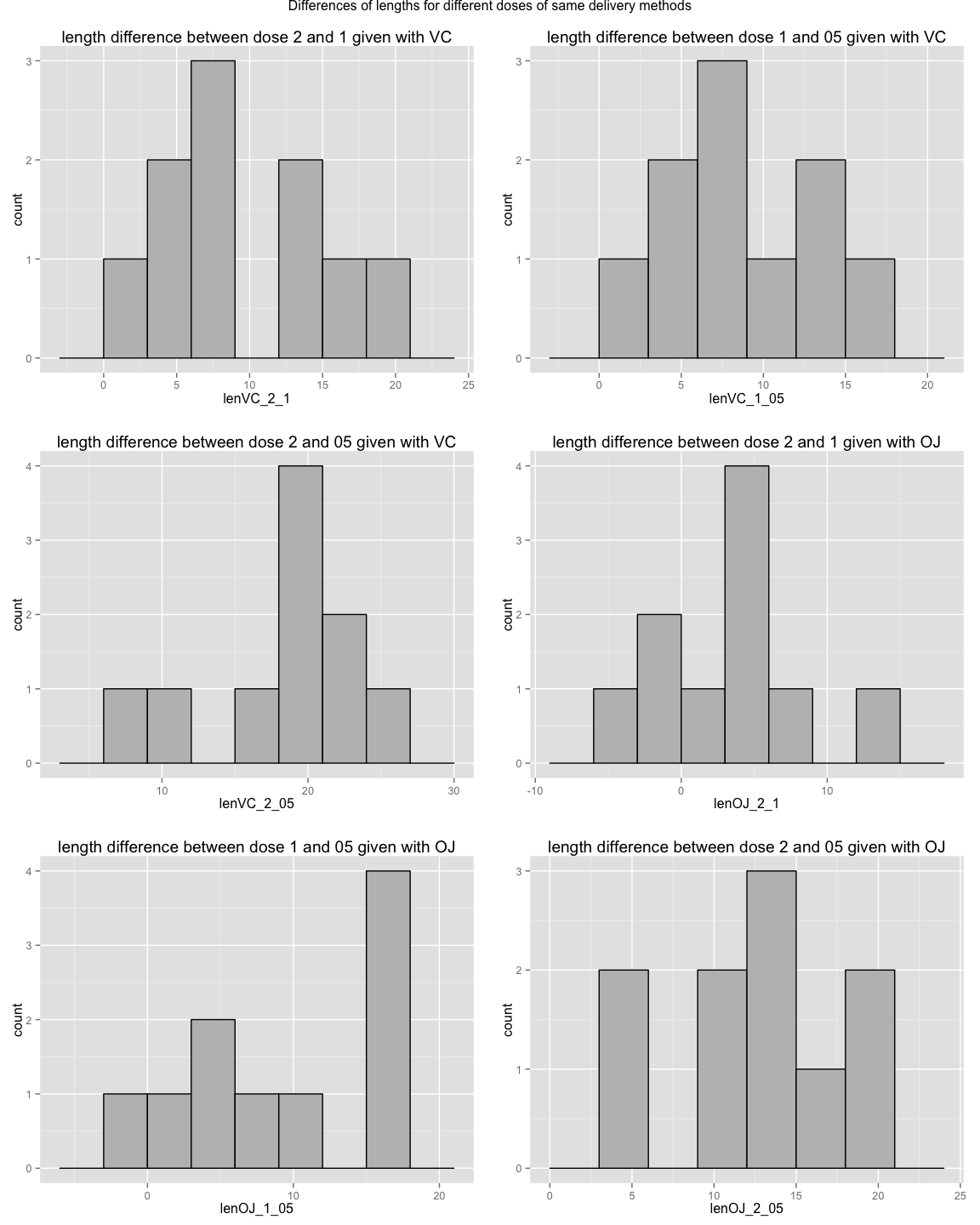
## lenVC\_1\_05 lenVC\_2\_1 lenVC\_2\_05 lenOJ\_1\_05 lenOJ\_2\_1 lenOJ\_2\_05  
## 1 12.3 7.1 19.4 4.5 5.8 10.3  
## 2 5.0 2.0 7.0 1.8 3.1 4.9  
## 3 7.9 18.7 26.6 6.0 -1.2 4.8  
## 4 11.5 8.2 19.7 16.7 -1.9 14.8  
## 5 16.1 3.9 20.0 5.5 4.8 10.3  
## 6 7.3 15.2 22.5 15.2 5.7 20.9  
## lenVC\_OJ\_05 lenVC\_OJ\_1 lenVC\_OJ\_2  
## 1 11.0 3.2 1.9  
## 2 10.0 6.8 7.9  
## 3 10.3 8.4 -11.5  
## 4 3.9 9.1 -1.0  
## 5 8.1 -2.5 -1.6  
## 6 0.0 7.9 -1.6

Now I will create a function to plot the differences between the same delivery methods. After that I will plot them adjusting the binwidth in order to approximate as best as I can normality.

#function for plotting the differences in the same delivery methods  
toothnormalplot <- function(mydf, bin, supp, doseup, dosedown) {  
 attach(mydf)  
 mytitle <- sprintf("length difference between dose %s and %s given with %s", doseup, dosedown, supp)  
 mydf$myx <- eval(parse(text = sprintf("len%s\_%s\_%s", supp, doseup, dosedown)))  
 g1 <- ggplot(data = mydf, aes(x = myx)) +  
 geom\_histogram(binwidth = bin, colour = "black", fill = "gray") +  
 ggtitle(label = mytitle) +  
 xlab(label = sprintf("len%s\_%s\_%s", supp, doseup, dosedown))  
 return(g1)  
 detach(mydf)  
}

g1 <- toothnormalplot(mydf = myToothGrowth\_diff, bin = 3, supp = "VC", doseup = "2", dosedown = "1")  
g2 <- toothnormalplot(mydf = myToothGrowth\_diff, bin = 3, supp = "VC", doseup = "1", dosedown = "05")  
g3 <- toothnormalplot(mydf = myToothGrowth\_diff, bin = 3, supp = "VC", doseup = "2", dosedown = "05")  
g4 <- toothnormalplot(mydf = myToothGrowth\_diff, bin = 3, supp = "OJ", doseup = "2", dosedown = "1")  
g5 <- toothnormalplot(mydf = myToothGrowth\_diff, bin = 3, supp = "OJ", doseup = "1", dosedown = "05")  
g6 <- toothnormalplot(mydf = myToothGrowth\_diff, bin = 3, supp = "OJ", doseup = "2", dosedown = "05")

grid.arrange(g1, g2, g3, g4, g5, g6, ncol = 2, main = "Differences of lengths for different doses of same delivery methods")



They are almost all quite normal, and considering that T test in quite robust to the normality assumtpions, we can be happy wity this results.

Now we should alo plot the differences for same dose but different delivery method. We will proceed as usual: first creating a function, second plotting.

#function for plotting the differences for same dose but different same delivery methods  
toothnormalplotsd <- function(mydf, bin, suppup, suppdown, dose) {  
 attach(mydf)  
 mytitle <- sprintf("length difference between delivery method %s and delivery method %s for dose %s", suppup, suppdown, dose)  
 mydf$myxsd <- eval(parse(text = sprintf("len%s\_%s\_%s", suppdown, suppup, dose)))  
 g1 <- ggplot(data = mydf, aes(x = myxsd)) +  
 geom\_histogram(binwidth = bin, colour = "black", fill = "gray") +  
 ggtitle(label = mytitle) +  
 xlab(label = sprintf("len%s\_%s\_%s", suppdown, suppup, dose))  
 return(g1)  
 detach(mydf)  
}

gsd1 <- toothnormalplotsd(myToothGrowth\_diff, 3, "OJ", "VC", "05")  
gsd2 <- toothnormalplotsd(myToothGrowth\_diff, 3, "OJ", "VC", "1")  
gsd3 <- toothnormalplotsd(myToothGrowth\_diff, 3, "OJ", "VC", "2")

grid.arrange(gsd1, gsd2, gsd3, ncol = 1, main = "Differences of lengths for same dose of different delivery method")

It seems that really only dose 0.5 is far away from normality. We can go through the test now.

### Hypothesis testing

I really need first a function to plot the curve of our test statistics and the region of rejection.

tplot <- function(dof, tstat, conflev, two = TRUE) {  
 myx <- seq(-4, 4, length=100)  
 hx <- dnorm(myx)  
 degf <- dof  
 mydf <- data.frame (myx, myy = dt(myx, degf))  
 pcum <- ifelse(two == TRUE, conflev+((1-conflev)/2), conflev)  
 tcrit <- qt(p = pcum, df = dof)  
 if (two == TRUE) {  
 g1 <- ggplot(mydf, aes(myx, myy)) +  
 geom\_line(size = 1.2, colour = "black") +  
 geom\_vline(xintercept = c(-tcrit, tcrit, tstat), colour = c("red", "red", "blue"), size = 0.9) +  
 ggtitle(paste("Hypothesis testing for t =", tstat ))  
 return(g1)  
 } else {  
 g2 <- ggplot(mydf, aes(myx, myy)) +  
 geom\_line(size = 1.2, colour = "black") +  
 geom\_vline(xintercept = c(tcrit, tstat), colour = c("red", "blue"), size = 0.9) +  
 ggtitle(paste("Hypothesis testing for t =", tstat ))  
 return(g2)  
 }  
}

Now let's perform the tests. As R provide a function to perform t test for paired samples, I am not going to use anymore the myToothGrowth\_diff dataframe, instead i will use the myToothGrowth3 one.

For every single difference vector I wnat to test, i will provide a plot and a comment. Looking at the plot remember that the red line represents the rejection region, while the blue one is the t statistics result of the test.

The alternative hypothesis will be *greater* per different doses of same delivery method, and *two.sided* for same dose of different delivery methods. The confidence level is .

#### Ascorbic acid: dose 2 and 0.5

* *null hypothesis*: mean of differences equal to 0
* *alternative hypothesis*: mean of differences greater than 0

tVC\_2\_05 <- t.test(x = myToothGrowth3$lenVC\_2, y = myToothGrowth3$lenVC\_05, alternative = "greater", paired = TRUE, conf.level = 0.95)  
tVC\_2\_05

##   
## Paired t-test  
##   
## data: myToothGrowth3$lenVC\_2 and myToothGrowth3$lenVC\_05  
## t = 9.7912, df = 9, p-value = 2.132e-06  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## 14.76006 Inf  
## sample estimates:  
## mean of the differences   
## 18.16

tplot(dof = 9, tstat = tVC\_2\_05$statistic, conflev = 0.95, two = FALSE)

The t statistic is incredibly high, with a p-value near to 0. Of course is not by chance that the mean of differences between the lengths derived by dose 2 and 0.5 is higher than 0, so **the null hypothesis is rejected**.

#### Ascorbic acid: dose 2 and 1

* *null hypothesis*: mean of differences equal to 0
* *alternative hypothesis*: mean of differences greater than 0

tVC\_2\_1 <- t.test(x = myToothGrowth3$lenVC\_2, y = myToothGrowth3$lenVC\_1, alternative = "greater", paired = TRUE, conf.level = 0.95)  
tVC\_2\_1

##   
## Paired t-test  
##   
## data: myToothGrowth3$lenVC\_2 and myToothGrowth3$lenVC\_1  
## t = 5.346, df = 9, p-value = 0.0002324  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## 6.157075 Inf  
## sample estimates:  
## mean of the differences   
## 9.37

tplot(dof = 9, tstat = tVC\_2\_1$statistic, conflev = 0.95, two = FALSE)

The t statistic, is higher than the critical value, with a p-value near to 0 as well, so **the null hypothesis is rejected**.

#### Ascorbic acid: dose 1 and 0.5

* *null hypothesis*: mean of differences equal to 0
* *alternative hypothesis*: mean of differences greater than 0

tVC\_1\_05 <- t.test(x = myToothGrowth3$lenVC\_1, y = myToothGrowth3$lenVC\_05, alternative = "greater", paired = TRUE, conf.level = 0.95)  
tVC\_1\_05

##   
## Paired t-test  
##   
## data: myToothGrowth3$lenVC\_1 and myToothGrowth3$lenVC\_05  
## t = 6.1364, df = 9, p-value = 8.576e-05  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## 6.16418 Inf  
## sample estimates:  
## mean of the differences   
## 8.79

tplot(dof = 9, tstat = tVC\_1\_05$statistic, conflev = 0.95, two = FALSE)

The t statistic is higher than the critical value, with a p-value near to 0 as well, so **the null hypothesis is rejected**.

So also a small difference in doses (only 0.5) if given with ascorbic acid is associated with a relevant growth of the teeth lengths.

#### Orange juice: dose 2 and 0.5

* *null hypothesis*: mean of differences equal to 0
* *alternative hypothesis*: mean of differences greater than 0

tOJ\_2\_05 <- t.test(x = myToothGrowth3$lenOJ\_2, y = myToothGrowth3$lenOJ\_05, alternative = "greater", paired = TRUE, conf.level = 0.95)  
tOJ\_2\_05

##   
## Paired t-test  
##   
## data: myToothGrowth3$lenOJ\_2 and myToothGrowth3$lenOJ\_05  
## t = 7.4919, df = 9, p-value = 1.862e-05  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## 9.690783 Inf  
## sample estimates:  
## mean of the differences   
## 12.83

tplot(dof = 9, tstat = tOJ\_2\_05$statistic, conflev = 0.95, two = FALSE)

The t statistic is incredibly high, with a p-value near to 0. Of course is not by chance that the mean of differences between the lengths derived by dose 2 and 0.5 is higher than 0, so **the null hypothesis is rejected**.

#### Orange juice: dose 2 and 1

* *null hypothesis*: mean of differences equal to 0
* *alternative hypothesis*: mean of differences greater than 0

tOJ\_2\_1 <- t.test(x = myToothGrowth3$lenOJ\_2, y = myToothGrowth3$lenOJ\_1, alternative = "greater", paired = TRUE, conf.level = 0.95)  
tOJ\_2\_1

##   
## Paired t-test  
##   
## data: myToothGrowth3$lenOJ\_2 and myToothGrowth3$lenOJ\_1  
## t = 1.9435, df = 9, p-value = 0.04192  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## 0.1908169 Inf  
## sample estimates:  
## mean of the differences   
## 3.36

tplot(dof = 9, tstat = tOJ\_2\_1$statistic, conflev = 0.95, two = FALSE)

Wow. For the first time the t statistic is close to the critical value, with a p-value a little bit smaller than the significance level. Anyway **the null hypothesis is rejected**, but with lower attained significance level. This is confirmed by the exploratory boxplot at the beginning of the report.

#### Orange juice: dose 1 and 0.5

* *null hypothesis*: mean of differences equal to 0
* *alternative hypothesis*: mean of differences greater than 0

tOJ\_1\_05 <- t.test(x = myToothGrowth3$lenOJ\_1, y = myToothGrowth3$lenOJ\_05, alternative = "greater", paired = TRUE, conf.level = 0.95)  
tOJ\_1\_05

##   
## Paired t-test  
##   
## data: myToothGrowth3$lenOJ\_1 and myToothGrowth3$lenOJ\_05  
## t = 4.1635, df = 9, p-value = 0.001218  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## 5.300497 Inf  
## sample estimates:  
## mean of the differences   
## 9.47

tplot(dof = 9, tstat = tOJ\_1\_05$statistic, conflev = 0.95, two = FALSE)

The t statistic is higher than the critical value, so **the null hypothesis is rejected**.

So also a small difference in doses (only 0.5) if given with orange juice is associated with a relevant growth of the teeth lengths.

#### Different delivery methods: dose 2

* *null hypothesis*: mean of differences equal to 0
* *alternative hypothesis*: mean of differences not equal to 0

t\_VC\_OJ\_2 <- t.test(x = myToothGrowth3$lenOJ\_2, y = myToothGrowth3$lenVC\_2, alternative = "two.sided", paired = TRUE, conf.level = 0.95)  
t\_VC\_OJ\_2

##   
## Paired t-test  
##   
## data: myToothGrowth3$lenOJ\_2 and myToothGrowth3$lenVC\_2  
## t = -0.0426, df = 9, p-value = 0.967  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -4.328976 4.168976  
## sample estimates:  
## mean of the differences   
## -0.08

tplot(dof = 9, tstat = t\_VC\_OJ\_2$statistic, conflev = 0.95, two = TRUE)

Huge p-value, t statistic close to 0. Of course **the null hypothesis is accepted**, and we are forced to say that the dose 2 given with different delivery method generates an average difference of teeth lengths equal to 0.

#### Different delivery methods: dose 1

* *null hypothesis*: mean of differences equal to 0
* *alternative hypothesis*: mean of differences not equal to 0

t\_VC\_OJ\_1 <- t.test(x = myToothGrowth3$lenOJ\_1, y = myToothGrowth3$lenVC\_1, alternative = "two.sided", paired = TRUE, conf.level = 0.95)  
t\_VC\_OJ\_1

##   
## Paired t-test  
##   
## data: myToothGrowth3$lenOJ\_1 and myToothGrowth3$lenVC\_1  
## t = 3.3721, df = 9, p-value = 0.008229  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 1.951911 9.908089  
## sample estimates:  
## mean of the differences   
## 5.93

tplot(dof = 9, tstat = t\_VC\_OJ\_1$statistic, conflev = 0.95, two = TRUE)

Wow. At dose 1 things change a lot, t statistic higher than critical value, so **the null hypothesis is rejected**. It seems that dose 1 given with orange juice produces a growth in lengths higher than the one given by ascorbic acid.

#### Different delivery methods: dose 0.5

* *null hypothesis*: mean of differences equal to 0
* *alternative hypothesis*: mean of differences not equal to 0

t\_VC\_OJ\_05 <- t.test(x = myToothGrowth3$lenOJ\_05, y = myToothGrowth3$lenVC\_05, alternative = "two.sided", paired = TRUE, conf.level = 0.95)  
t\_VC\_OJ\_05

##   
## Paired t-test  
##   
## data: myToothGrowth3$lenOJ\_05 and myToothGrowth3$lenVC\_05  
## t = 2.9791, df = 9, p-value = 0.01547  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 1.263458 9.236542  
## sample estimates:  
## mean of the differences   
## 5.25

tplot(dof = 9, tstat = t\_VC\_OJ\_05$statistic, conflev = 0.95, two = TRUE)

Alos for dose 0.5 t statistic is higher than critical value, so **the null hypothesis is rejected**.

## Conclusions

ascorbic acid da più potere alla dose di orange juice.