Investigation of Tooth Growth Data in R

Julian Hatwell

knitr::opts\_chunk$set(warning=FALSE, message=FALSE)

## Summary

This document describes a statistical inference exercise relating to the effect of supplements on tooth growth in guinea pigs. The tools of statistical inference and the R programming environment will be used to explore the ToothGrowth data set.

Tooth growth in guinea pigs can be said to increase with increasing dose of a supplement, either orange juice or vitamin C with >95% condidence. However, it is not possible to determine from this data whether orange juice or vitamin C produce a greater effect.

### Exploratory Data Analysis

Reading the help file for ToothGrowth provides the following information:

"The response is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid)."

The following code runs a quick summary to double check the information:

summary(cbind(ToothGrowth, factor(ToothGrowth$dose)))

## len supp dose factor(ToothGrowth$dose)  
## Min. : 4.20 OJ:30 Min. :0.500 0.5:20   
## 1st Qu.:13.07 VC:30 1st Qu.:0.500 1 :20   
## Median :19.25 Median :1.000 2 :20   
## Mean :18.81 Mean :1.167   
## 3rd Qu.:25.27 3rd Qu.:2.000   
## Max. :33.90 Max. :2.000

There are 60 observations as expected over 10 animals on 2 supplements each given at 3 different doses. There are no NA values.

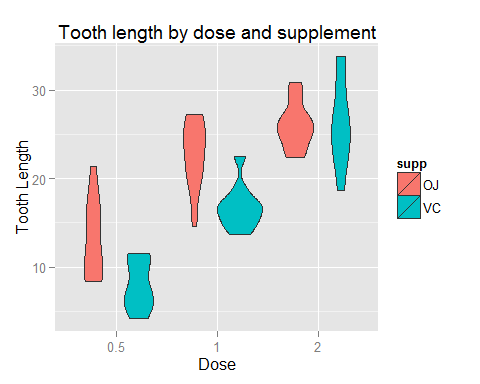
It's also useful to know from the description that the data is collected from always the same 10 specimens. It is reasonable to infer that any change in tooth growth is a response to the supplement and not a difference between individual animals.

However, a flaw is that the dataset does not identify the individual animals and a search of online references does not give any confirmation or otherwise if the set is ordered. Therefore paired t-tests cannot be relied upon, although paired and unpaired tests could be run for comparison.

The following code produces a violin plot of tooth length alone, as a first look at variability of the data.

A violin plot provides visual information that is something between a box plot and a density plot:

library(ggplot2)  
  
g <- ggplot(ToothGrowth, aes(x=factor(dose), y=len, fill=supp))   
g <- g + geom\_violin()  
g <- g + labs(title = "Tooth length by dose and supplement"  
 , y = "Tooth Length"  
 , x = "Dose")  
g <- g + scale\_color\_discrete("Supplement")  
g



#### Summarising the analysis so far:

From the above graph, there appear to be two factors playing out on the tooth growth.

1. There appears to be a clear trend of increasing tooth growth with increasing supplement dose.
2. Less clear is whether or not the specific supplement makes a difference (pure vitamin C vs orange juice).

The effect of dose seems to be intuitive from the graph. Nevertheless, this can be quantified and will be tested properly to get a decent benchmark for drawing conclusions.

The following code separates the data into samples according to dose (regardless of supplement) and runs t-tests on the samples. This is done in two steps because the t-test function requires factors to have no more than 2 levels.

half <- ToothGrowth$len[ToothGrowth$dose == 0.5]  
one <- ToothGrowth$len[ToothGrowth$dose == 1]  
two <- ToothGrowth$len[ToothGrowth$dose == 2]  
  
tp1 <- t.test(half, one, paired = TRUE)  
tu1 <- t.test(half, one, paired = FALSE)  
  
tp2 <- t.test(one, two, paired = TRUE)  
tu2 <- t.test(one, two, paired = FALSE)

The confidence intervals for dose 0.5 vs dose 1.0, for paired and unpaired tests, respectively are:

tp1$conf.int

## [1] -11.872879 -6.387121  
## attr(,"conf.level")  
## [1] 0.95

tu1$conf.int

## [1] -11.983781 -6.276219  
## attr(,"conf.level")  
## [1] 0.95

The confidence intervals for dose 1 vs dose 2, for paired and unpaired tests, respectively are:

tp2$conf.int

## [1] -9.258186 -3.471814  
## attr(,"conf.level")  
## [1] 0.95

tu2$conf.int

## [1] -8.996481 -3.733519  
## attr(,"conf.level")  
## [1] 0.95

The following code runs a t-test on the samples according to supplement (regardless of dose).

tp <- t.test(len ~ supp, paired = TRUE, data = ToothGrowth)  
tu <- t.test(len ~ supp, paired = FALSE, data = ToothGrowth)

The confidence intervals for OJ vs VC, for paired and unpaired tests, respectively are:

tp$conf.int

## [1] 1.408659 5.991341  
## attr(,"conf.level")  
## [1] 0.95

tu$conf.int

## [1] -0.1710156 7.5710156  
## attr(,"conf.level")  
## [1] 0.95

## Conclusions

The data clearly show that increasing the dose of supplement increases tooth growth and this is the case whether or not we assume the data sets were paired.

It is not certain that pure vitamin C is better or worse than orange juice because the unpaired test confidence interval contains 0, meaning they might be the same. The null hypothesis is not rejected.

The results from the paired test cannot be relied upon because there is no indication from the data or from external sources that the data a ordered and can be paired.

### Assumptions

t-tests such as shown in the above demonstration rely on a number of assumptions:

1. The data should be symmetrically distributed. This is not the case in the guinea pig data set as can be seen from the violin plots. Other tests that the t-test may be more appropriate in these cases, but t-test was selected because it is ideally suited to small samples.
2. The pairing relationship between samples must be known. This is not the case im the guinea pig data, and so the accuracy cannot be correctly determined.
3. The equality of variance of the samples under comparison must be known (or shall be assumed different). This was beyond the scope of the above demonstration.