Statistical Inference Course Project: Inferential Data Analysis

Philip Bulsink

2017-01-28

Introduction

This report will investigate the `R` dataset `ToothGrowth`. The data will be summarized and relationships between tooth length and supplement or dosage will be tested using confidence intervals and hypothesis tests, as described in the Coursera Data Science course on Statistical Inference.

Data Exploration

The `ToothGrowth` dataset contains three columns of data. The column named `len` lists tooth length, the column `supp` describes the supplement type (either VC or OJ), and the column `dose` gives the dosage of the supplement in milligrams per day.

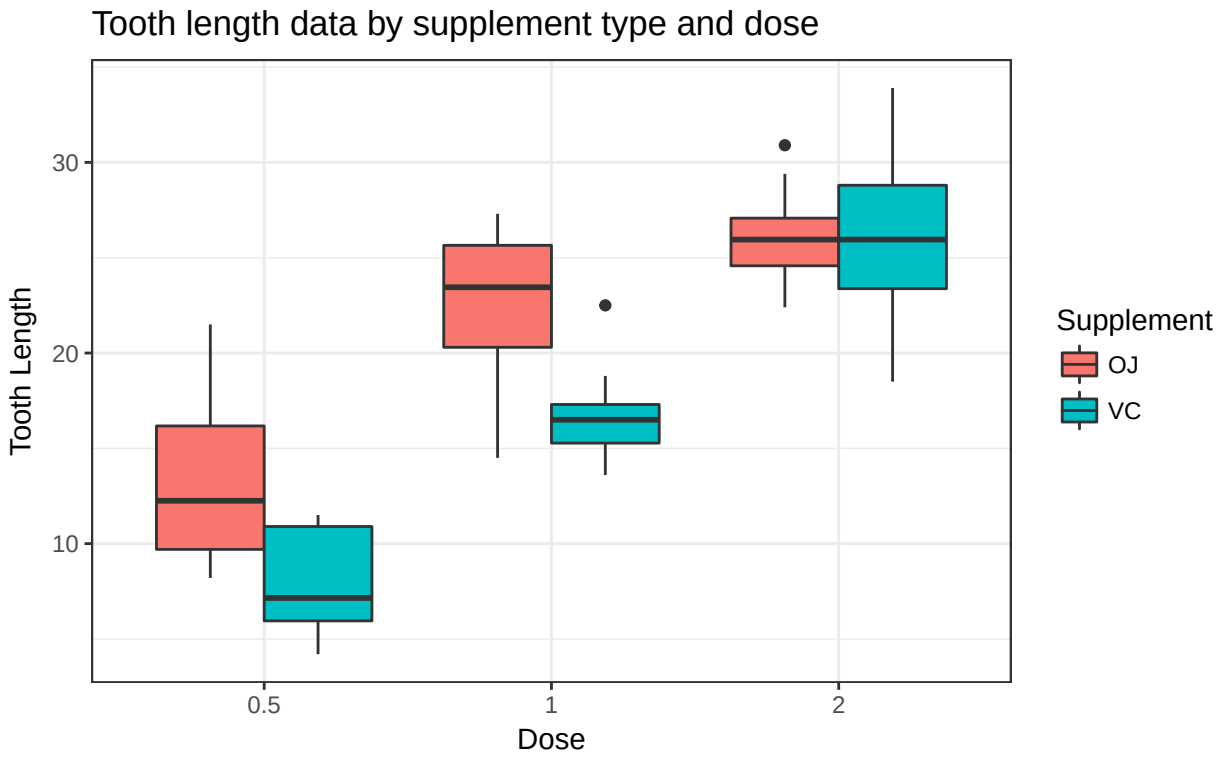
```
data("ToothGrowth")
summary(ToothGrowth)
```

```
##       len       supp         dose
##  Min.   : 4.2   OJ:30   Min.   :0.50
##  1st Qu.:13.1   VC:30   1st Qu.:0.50
##  Median :19.2           Median :1.00
##  Mean   :18.8           Mean   :1.17
##  3rd Qu.:25.3           3rd Qu.:2.00
##  Max.   :33.9           Max.   :2.00
```

We can see that there are two supplements, each given at a range of 0.5 to 2 mg/day, resulting in a range of tooth lengths.

We'll plot the data with box plots, showing the relationship between dose and toothlength, broken up by supplement type.

```
ggplot(ToothGrowth, aes(as.factor(dose), len, fill = supp)) +
    geom_boxplot() +
    ggtitle("Tooth length data by supplement type and dose")+
    xlab("Dose")+
    ylab("Tooth Length") +
    guides(fill=guide_legend(title="Supplement")) +
    theme_bw()
```



There's obviously some relationship there, let's test for them.

Testing

There are two things we can test in this data. We can test to see if the choice of supplement impacts the tooth length, and we can test if dose amount impacts tooth data.

Confidence Interval Testing of Dose Size

We'll start by testing for significance of dose (irrespective of supplement type). We'll use the `R` function `t.test()` to test the significance between two groups at a time. Note that this data is unpaired, no data point is related to any other point in any way. Note that the data will be broken up into three groups to simplify the coding

```
g1 <- ToothGrowth[ToothGrowth$dose == 0.5,]
g2 <- ToothGrowth[ToothGrowth$dose == 1, ]
g3 <- ToothGrowth[ToothGrowth$dose == 2, ]

tt1<-t.test(g2$len, g1$len)$conf[1:2]
tt2<-t.test(g3$len, g2$len)$conf[1:2]
tt3<-t.test(g3$len, g1$len)$conf[1:2]
```

There are three tests performed, comparing 0.5 to 1 mg/day, 1 to 2 mg/day, and 0.5 to 2 mg/day. The null hypothesis states that there is no difference between means of tooth length between dosing groups. When we calculate the t.test, we are provided with the 95% confidence interval for the difference. If that interval includes 0, then we can't reject the null hypothesis, or, we can't statistically say that the dosing makes a difference.

The results of the tests are:

	Low CI	High CI
0.5 to 1 mg	6.28	9.0
1 to 2 mg	11.98	12.8
0.5 to 2 mg	3.73	18.2

Thus, we can see that at no time does the confidence interval include 0, each increase in dosing increases tooth length, with statistically significant confidence.

Permutation Testing of Supplement

One way of testing to see if two groups are statistically different is to test the permutation of their results. That is, to randomly assign group labels to each values, and test for confidence. We'll test each supplement group at each dosing amount, to see if one supplement is better than another when given at a certain dose.

This code splits lengths from each dosing level, calculates the difference in means, then performs 10,000 tests of permuting the group label over the data.

```
len1<-g1$len; len2<-g2$len; len3<-g3$len
group<-as.character(g1$supp) #same for g2 and g3
testStat<-function(w, g) mean(w[g=="OJ"])-mean(w[g=="VC"])
observed1<-testStat(len1, group)
observed2<-testStat(len2, group)
observed3<-testStat(len3, group)
perm1<-sapply(1:10000, function(i) testStat(len1, sample(group)))
perm2<-sapply(1:10000, function(i) testStat(len2, sample(group)))
perm3<-sapply(1:10000, function(i) testStat(len3, sample(group)))
```

We compare these permuted results with the original difference in the means by looking at how often the permuted groups give a larger mean difference than the original group. This gives us an estimate of our p value for rejecting the null hypothesis, that is, how confident we are to say that the difference in supplements is statistically significant.

```
mean(perm1>observed1)
```

```
## [1] 0.002
```

This provides us with the confidence interval to say that the mean of the `OJ` supplement is higher than that of the `VC` supplement, when given 0.5 mg/day dose. The difference of those means is 5.25.

Similarly, we find that we have a p value of 0.0012 for our 1 mg/day dosage (with an average tooth length difference of 5.93), and a p value of 0.513 for 2 mg/day dosage (with an average difference of -0.08).

Summary

We have identified the statistical differences between supplements and dosages in the `R` dataset `ToothGrowth`. Each dosage level is statistically different from the others. The differences between supplements is statistically significant for dosage levels of 0.5 and 1 mg/day, but not for 2 mg/day.