ToothGrowth Data Analysis

Benjamin Phillips

Course project for the course Inferential Statistics, part of the Coursera John Hopkins data-science specialisation. Project and code information can be found on my github account

Data

The Effect of Vitamin C on Tooth Growth in Guinea Pigs

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). len numeric Tooth length supp factor Supplement type (VC or OJ). dose numeric Dose in milligrams/day

Load packages

```
library(ggplot2)
library(dplyr)
```

Have a look at the data

```
data("ToothGrowth")
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 ...
```

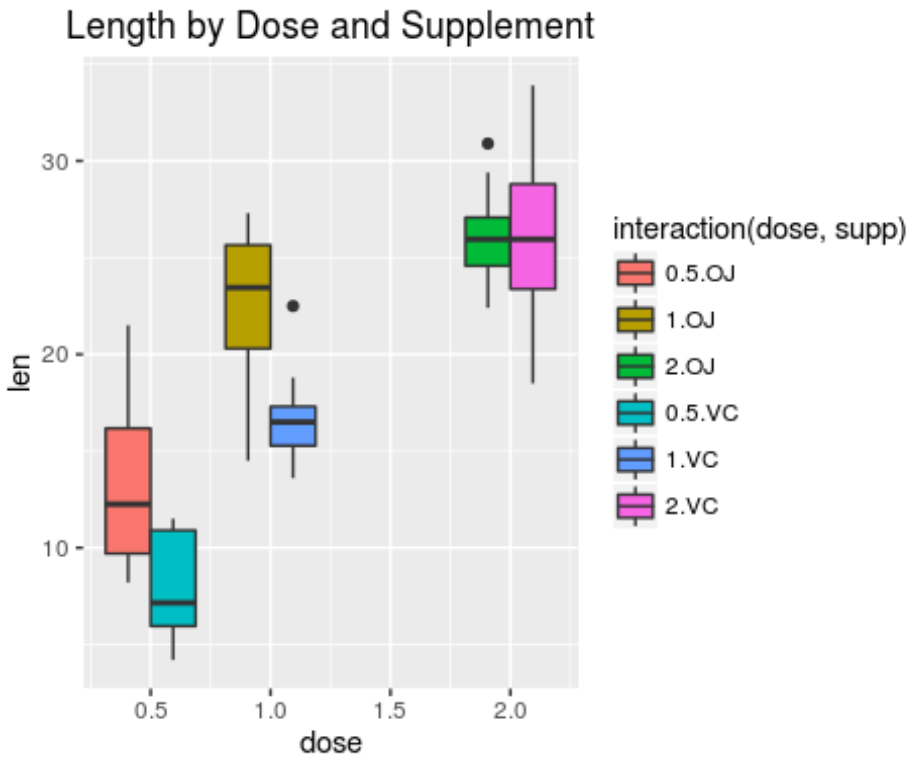
```
table(ToothGrowth$supp)
```

```
##
## OJ VC
## 30 30
```

```
table(ToothGrowth$dose)
```

```
##
## 0.5   1   2
##  20  20  20
```

```
g1 <- ggplot(data = ToothGrowth, aes(dose, len, fill = interaction(dose, supp))) +
   geom_boxplot() +
   labs(title = "Length by Dose and Supplement")
print(g1)
```



There are two types of supplements (supp) and three types of doses (dose)

Lets first consider the effect of dose. H0 is that dose has no effect on length and Ha is that is does have an effect. We will need to test for both supps seperately. alpha = 0.05

```
supps <- c("OJ", "VC")
doses <- c(0.5, 1, 2)

doses_p.values <- numeric(length = length(supps)*length(doses))
i <- 1
for(s in supps){
   for(d1 in doses){
      for(d2 in doses[doses>d1]){
         doses_p.values[i] <- t.test(ToothGrowth %>% filter(supp==s,
dose==d1) %>% select(len),
                                     ToothGrowth %>% filter(supp==s, dose==d2) %>
% select(len)
                                     )$p.value
         i <- i+1
      }
   }
}
```

Due to multiple testing, we should control for False Discovery Rate (FDR, Benjamin Hochberg)

```
p.adjust(doses_p.values, method = "BH")
```

```
## [1] 1.098672e-04 2.647568e-06 3.919514e-02 2.043305e-06 2.808946e-07
## [6] 1.098672e-04
```

All of these p-values are below alpha, we can reject the null hypothesis that dose has no effect.

Now look at supplements. H0 is that there is no difference between supplements and Ha is that there is a difference between them.

```
supps_p.values <- numeric(length = length(doses))
i <- 1
for(d in doses){
   supps_p.values[i] <- t.test(ToothGrowth %>% filter(supp=="OJ", dose==d) %>% select(len),
          ToothGrowth %>% filter(supp=="VC", dose==d) %>% select(len)
          )$p.value
   i <- i+1
}
```

For multiple testing, we should control for False Discovery Rate (FDR, Benjamin Hochberg)

```
p.adjust(supps_p.values, method = "BH")
```

```
## [1] 0.009537910 0.003115128 0.963851589
```

For doses 0.5 and 1.0, it appears that supplement does have an effect, but for dose 2.0 we have a p-value of 0.95 and cannot reject the null hypothesis that the supplement has no effect.

Discussion

To perform the t.tests we assumed that the data was normally distributed. Two sample t.tests also assume both samples have the same variance, though it is robust to the presence of unequal variances.

For the first set of t.tests (6 comparisons), we recevied very small p-values which makes the null hypothesis (that dose has no effect) very unlikely, and we can reject it.

For the second set of t.tests (3 comparisons), small p-values for doses 0.5 and 1.0 indicate that it is very unlikey that there is no difference between orange juice (OJ) or ascorbic acid (VC) but for doses of 2.0 we don't have enough evidence to reject the null hypothesis.