Part 2: Basic Inferential Data Analysis

Maurício Collaça

October 24 2017

Overview

An analysis of the ToothGrowth data from the R datasets through exploratory data analysis and comparison of the tooth growth by supp and dose using multiple hypothesis testing. The ToothGrowth data is about the Effect of Vitamin C on tooth growth in pigs. The dependent variable is the length found in 60 pigs. Each animal received one of dose levels of treatment (0.5, 1, and 2 mg/day) by one of two delivery methods: orange juice or vitamin C.

Exploratory data analyses

ToothGrowth is a data frame with 60 observations on 3 variables:

```
library(ggplot2); library(dplyr)
data("ToothGrowth")
str(ToothGrowth)
```

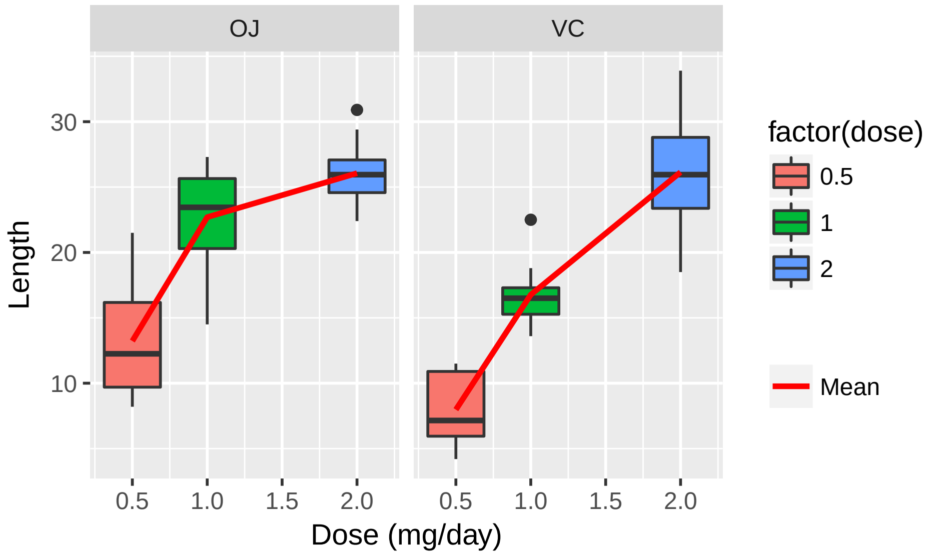
```
'data.frame':   60 obs. of  3 variables:
 $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
 $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
 $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

Some assumptions about the data:

- It's a randomized trial or AB test with no placebo group.
- Length is a continuous variable, supplement is a categorical nominal variable and dose is categorical ordinal variable.
- There is an interaction between Supplement and Dose.
- There are no missing values or unexpected supplements or dose levels.

The following boxplot shows tooth length (y) by dose (x) split by supplement. The line shows mean variation between groups. It increases as doses increase, but at different rates between supplements.

```
data <- ToothGrowth %>% group_by(supp, dose) %>% mutate(mean = mean(len))
ggplot(data, aes(x=dose, y=len, group=dose, fill=factor(dose))) + facet_grid(~ supp) +
    geom_boxplot() + geom_line(aes(y=mean, group=NA, color="Mean"), size=1) +
    scale_color_manual(NULL, values = c(Mean="red")) + xlab("Dose (mg/day)") + ylab("Length")
```



It suggests null hypothesis that group means are all equal is not true and it must pass through formal multiple hypothesis test.

Summary of the data

Based on the assumptions made so far, a basic summary of the data that supports the exploratory plots and a multiple hypothesis test design contains the following statistics broken down by supplement type and dose:

```
ToothGrowth %>% group_by(supp, dose) %>%
    summarize(count=n(), mean=mean(len), variance=var(len)) %>% knitr::kable()
```

supp	dose	count	mean	variance
OJ	0.5	10	13.23	19.889000
OJ	1.0	10	22.70	15.295556
OJ	2.0	10	26.06	7.049333
VC	0.5	10	7.98	7.544000
VC	1.0	10	16.77	6.326778
VC	2.0	10	26.14	23.018222

Tooth growth comparison

Hypothesis formulation

The null hypothesis is that all of the means of supplement doses are equal, i.e., neither the supplement type being orange juice or vitamin C or their doses being 0.5, 1 or 2 mg/day would make any difference in the tooth growth.

$$H_0 : \mu_{OJ,0.5} = \mu_{OJ,1} = \mu_{OJ,2} = \mu_{VC,0.5} = \mu_{VC,1} = \mu_{VC,2}$$

The alternative hypothesis is at least one of the means of supplement doses are different from the others, i.e., either the supplement type or a specific dose would make a difference in the tooth growth.

$$H_1 : \mu_{OJ,0.5} \neq \mu_{OJ,1} \neq \mu_{OJ,2} \neq \mu_{VC,0.5} \neq \mu_{VC,1} \neq \mu_{VC,2}$$

The relevant hypothesis test is a t-test for independent groups based on this assumptions:

- The sample size is relatively small: 10 for each group
- The samples are independent

The tests to be performed are a combination of existing 6 levels of supplement and dose, taken two at a time:

```
ToothGrowth$group <- paste0(ToothGrowth$supp, ToothGrowth$dose)
glimpse(tests <- as.data.frame(t(combn(unique(ToothGrowth$group), 2))))
```

```
Observations: 15
Variables: 2
$ V1 <fctr> VC0.5, VC0.5, VC0.5, VC0.5, VC0.5, VC1, VC1, VC1, VC1, VC2...
$ V2 <fctr> VC1, VC2, OJ0.5, OJ1, OJ2, VC2, OJ0.5, OJ1, OJ2, OJ0.5, OJ...
```

Multiple hypothesis t-tests

The following idiom performs the multiple t-test for the 15 combinations.

```
tests <- tests %>%
    group_by(V1,V2) %>%
    mutate(pval = t.test(subset(ToothGrowth, group==V1, "len"),
                         subset(ToothGrowth, group==V2, "len"))$p.value) %>%
    ungroup() %>% arrange(pval)
```

False discoveries control

As there are multiple hypothesis, tests must be corrected to avoid false positives or discoveries. The approach used is to calculate adjusted p-values with these methods:

- Family-Wise Error Rate (FWER) using "bonferroni" correction
- False Discovery Rate (FDR) using "BH" and "BY" corrections

The following idiom calculates adjusted p-values by the three methods, each with different null hypothesis result. The true discoveries are the ones whose H_0 is `FALSE`.

```r
tests <- tests %>%
    mutate(uncorrected.H0 = pval >= 0.05, BF.pval = p.adjust(pval, method = "bonferroni"),
           BF.H0 = BF.pval >= 0.05, BH.pval = p.adjust(pval, method = "BH"),
           BH.H0 = BH.pval >= 0.05, BY.pval = p.adjust(pval, method = "BY"),
           BY.H0 = BY.pval >= 0.05)
tests %>% knitr::kable(col.names = c("$\\mu_1$", "$\\mu_2$", "p-value", "Uncorrected $H_0$?",
                                      "Bonferroni", "$H_0$?", "BH", "$H_0$?", "BY", "$H_0$?"))
```

μ_1	μ_2	p-value	Uncorrected H_0?	Bonferroni	H_0?	BH	H_0?	BY	H_0?
VC0.5	OJ2	0.0000000	FALSE	0.0000000	FALSE	0.0000000	FALSE	0.0000000	FALSE
VC0.5	OJ1	0.0000000	FALSE	0.0000005	FALSE	0.0000002	FALSE	0.0000008	FALSE
VC0.5	VC2	0.0000000	FALSE	0.0000007	FALSE	0.0000002	FALSE	0.0000008	FALSE
VC1	OJ2	0.0000002	FALSE	0.0000035	FALSE	0.0000009	FALSE	0.0000029	FALSE
VC0.5	VC1	0.0000007	FALSE	0.0000102	FALSE	0.0000020	FALSE	0.0000068	FALSE
OJ0.5	OJ2	0.0000013	FALSE	0.0000199	FALSE	0.0000033	FALSE	0.0000110	FALSE
VC2	OJ0.5	0.0000072	FALSE	0.0001079	FALSE	0.0000154	FALSE	0.0000512	FALSE
OJ0.5	OJ1	0.0000878	FALSE	0.0013177	FALSE	0.0001526	FALSE	0.0005063	FALSE
VC1	VC2	0.0000916	FALSE	0.0013733	FALSE	0.0001526	FALSE	0.0005063	FALSE
VC1	OJ1	0.0010384	FALSE	0.0155756	FALSE	0.0015576	FALSE	0.0051684	FALSE
VC0.5	OJ0.5	0.0063586	FALSE	0.0953791	TRUE	0.0086708	FALSE	0.0287718	FALSE
OJ1	OJ2	0.0391951	FALSE	0.5879271	TRUE	0.0489939	FALSE	0.1625731	TRUE
VC1	OJ0.5	0.0460103	FALSE	0.6901550	TRUE	0.0530888	TRUE	0.1761609	TRUE
VC2	OJ1	0.0965261	TRUE	1.0000000	TRUE	0.1034208	TRUE	0.3431741	TRUE
VC2	OJ2	0.9638516	TRUE	1.0000000	TRUE	0.9638516	TRUE	1.0000000	TRUE

Conclusions

The exploration data analysis suggested a mean variation between groups that increases as doses increase, but at different rates between supplements. The null hypothesis that group means are all equal is true which leads to a multiple hypothesis t-test for independent groups. As a multiple tests must be corrected in order to avoid false discoveries, it's been decide to apply three types of corrections and compare them.

Without any error correction, it's found 13 discoveries.

Boferroni adjustment (FWER) found 10 discoveries which means that 3 were false discoveries.

BH adjustment (FDR) found 12 discoveries which means that 1 were false discoveries.

BY adjustment (FDR) found 11 discoveries which means that 2 were false discoveries.