Analysis of tooth growth dataset - Statistical Inference Course Project

Armando Güereca

September 24th, 2015

Overview

We're going to analyze the ToothGrowth data in the R datasets package, our analysis will cover: 1. Load the ToothGrowth data and perform some basic exploratory data analyses 2. Provide a basic summary of the data. 3. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use 4. the techniques from class, even if there's other approaches worth considering) 5. State your conclusions and the assumptions needed for your conclusions.

Exploratory data analysis

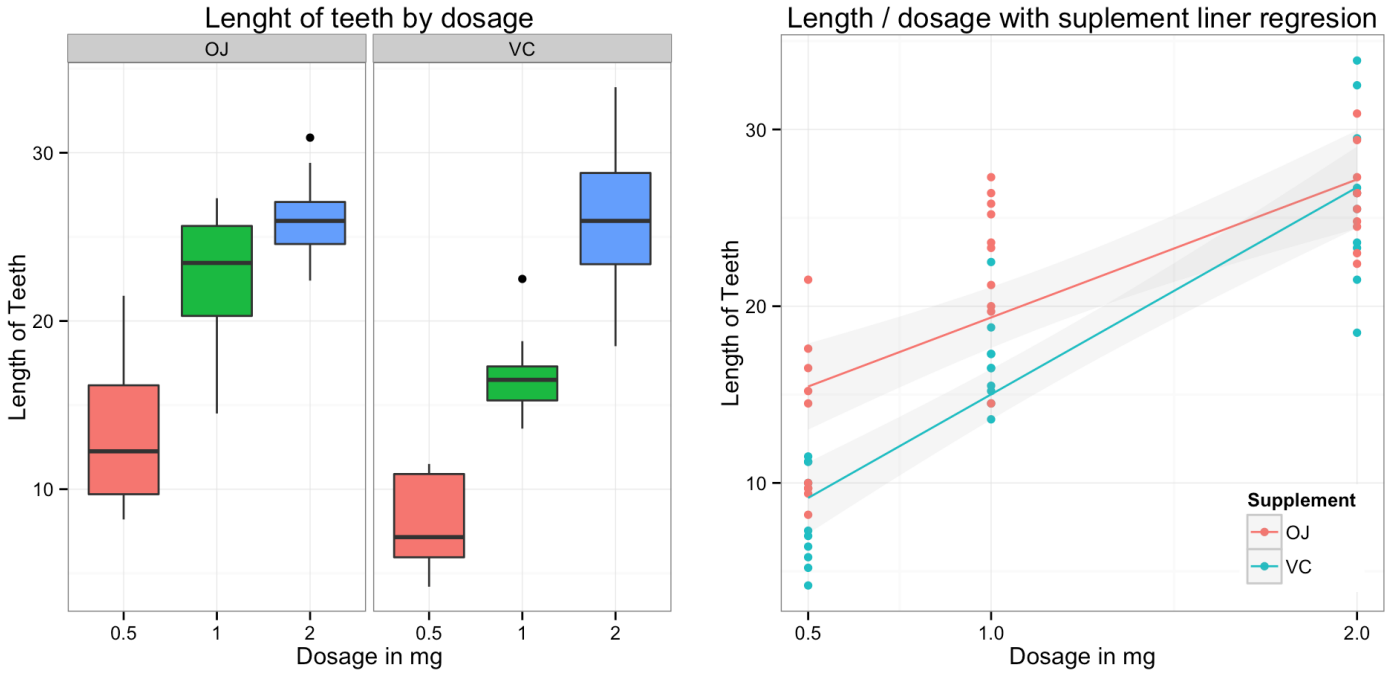
We start our analysis by loading and summarizing the ToothGrowth dataset which consists of the response in the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (OJ for orange juice, or VC for Vitamin C):

```
# Load our source dataset
tg_data <- ToothGrowth
# Summarize by suplement and dosage
aux <- tg_data %>% group_by(supp, dose) %>%
    summarise(Mean=mean(len), Min=min(len), Max=max(len), "Std. Dev"=sd(len))
# Format data aggregating by dosage
aux2 <- dcast(melt(aux, id.vars=c("supp", "dose")),
              dose ~ supp + variable, fun.aggregate=sum)
aux2 <- round(aux2,2)
names(aux2)[1] <- "Dosage"
kable(aux2, format = "markdown")
```

Dosage	OJ_Mean	OJ_Min	OJ_Max	OJ_Std. Dev	VC_Mean	VC_Min	VC_Max	VC_Std. Dev
0.5	13.23	8.2	21.5	4.46	7.98	4.2	11.5	2.75
1.0	22.70	14.5	27.3	3.91	16.77	13.6	22.5	2.52
2.0	26.06	22.4	30.9	2.66	26.14	18.5	33.9	4.80

The table above shows summary statistics for the Tooth Growth data. It suggest that orange juice is more effective at lower dosage but also that VC supplements are similarly effective at the 2mg dose. It also suggests that Vitamin C has less variability for lower doses and more for the higher.

Lets further explore the relation of length of teeth and dosage by plotting their statistics:



This figures shows the impact of the supplement and the dosage on the length of teeth of Guinea Pigs. The left plot shows the distributions (in boxplots) for the different combinations of supplement and dosage, the fact that the boxes do not overlap might be strong evidence that their medians are different and if we assume normality (which is logical for biological patterns like growth), the different means are expected as the median is the same as the mean for normal distributions. The right plot shows a scatter plot of the data with linear regression fits of teeth length as a function of dosage for each supplement.

Both figures are strong evidence that higher dosages appears to be related to longer teeth (however there appears there may be some diminishing returns for larger doses of orange juice). It is also possible to see that for dosage of 0.5 and 1 mg, orange juice appears to be far superior, whereas Vitamin C could potentially lead to longer teeth at the 2mg dose.

Compare Peformance and Dosage

- Is Orange Juice more effective than Vitamin C for similar doses?

```
# T-Testing with and without assumed equality on variance
tt1 <- t.test(len~supp, paired=F, var.equal=T, data=tg_data, alternative="greater")
tt2 <- t.test(len~supp, paired=F, var.equal=F, data=tg_data, alternative="greater")
# P-Values
c(round(tt1$p.value, 4), round(tt2$p.value, 4))
# P-Values with Bonferroni adjustment
c(round(tt1$p.value*50, 4), round(tt2$p.value*50, 4))
# Confidence intervals:
c(round(tt1$conf.int, 4), round(tt2$conf.int, 4))
```

```
## [1] 0.0302 0.0303
## [1] 1.5098 1.5159
## [1] 0.4708    Inf 0.4683    Inf
```

By the low p-value of the a T-tests (either with equal or unequal variance) we conclude (with 95% confidence) that orange juice had a greater impact on tooth growth than Vitamin C, also as the confidence intervals do not contain zero we rejected the null hypothesis. However if we use the Bonferroni adjusted values we can't conclude that there is any difference in the 2 supplements

- Does Dosage Matter?

```
# We will reuse this columns on the next  tables
common_cols = c(low_lim=numeric(), up_lim=numeric(), pval=numeric(), variance_equal=log
ical(),
                test=character(), stringsAsFactors=F)

# Define data frame to hold our analysis's results
res_dos = do.call(data.frame,
                  as.list(supplement=character(), dose_1=numeric(), dose_2=numeric(), c
ommon_cols))
k = 1
# Perform T-Test on subsets of each dose and suplement
# Each test with and without assumed equality on variance and for both alternative hypo
thesis
for (i in levels(tg_data$supp)) {
  for (j in unique(tg_data$dose)) {
    for (test in c("greater", "less")) {
      for (variance in c(T, F)) {
        data_sub_set <- subset(tg_data, supp==i & dose!=j)
        temp <- t.test(len~dose, paired=F, var.equal=variance, data=data_sub_set, alter
native=test)
        res_dos[k, "supplement"] <- i
        res_dos[k,"dose_1"] <- unique(data_sub_set$dose)[1]
        res_dos[k, "dose_2"] <- unique(data_sub_set$dose)[2]
        res_dos[k,"low_lim"] <- temp$conf.int[1]
        res_dos[k, "up_lim"] <- temp$conf.int[2]
        res_dos[k,"pval"] <- temp$p.value
        res_dos[k, "variance_equal"] <- variance
        res_dos[k,"test"] <- test
        k <- k+1
      }
    }
  }
}
# P-Values with Bonferroni adjustment
res_dos$adj_pval <- res_dos$pval*50
# P-value interpretation
res_dos$outcome <- ifelse(res_dos$pval<=0.05, "Difference", "No Difference")
# Formating
res_dos[c(4:6,9)] <- round(res_dos[c(4:6,9)], 3)
res_dos_diff <- subset(res_dos, outcome=="Difference")
res_dos_diff <- res_dos_diff[,c(1:6,9,7,8,10)]
kable(res_dos_diff, format = "markdown")
```

	supplement	dose_1	dose_2	low_lim	up_lim	pval	adj_pval	variance_equal	test	outcome
3	OJ	1.0	2	-Inf	-0.768	0.019	0.934	TRUE	less	Difference

4	OJ	1.0	2	-Inf	-0.749	0.020	0.980	FALSE	less	Difference
7	OJ	0.5	2	-Inf	-9.984	0.000	0.000	TRUE	less	Difference
8	OJ	0.5	2	-Inf	-9.948	0.000	0.000	FALSE	less	Difference
11	OJ	0.5	1	-Inf	-6.217	0.000	0.002	TRUE	less	Difference
12	OJ	0.5	1	-Inf	-6.214	0.000	0.002	FALSE	less	Difference
15	VC	1.0	2	-Inf	-6.399	0.000	0.001	TRUE	less	Difference
16	VC	1.0	2	-Inf	-6.347	0.000	0.002	FALSE	less	Difference
19	VC	0.5	2	-Inf	-15.129	0.000	0.000	TRUE	less	Difference
20	VC	0.5	2	-Inf	-15.086	0.000	0.000	FALSE	less	Difference
23	VC	0.5	1	-Inf	-6.748	0.000	0.000	TRUE	less	Difference
24	VC	0.5	1	-Inf	-6.747	0.000	0.000	FALSE	less	Difference

The table above shows the dosage comparisons for each supplement for which a statistical difference of at least 95% was observed. It shows that higher dosages are more effective that lower dosages for either supplement. However, when we use the Bonferroni adjustment for p-values (using n=50 since in general we conducted 50 comparisons), we can't conclude anymore that the 2mg dosage for orange juice is more effective than the 1mg dosage.

- Which Supplement is Better at a Given Dosage?

```
# Define data frame to hold our analysis's results, use same columns as before
res_sup_dos = do.call(data.frame,
                      as.list(dose=numeric(), supplement1=character(), supplement2=char
acter(),
                              common_cols))
k=1
# Perform T-Test on subsets of each unique dose and suplement
# Each test with and without assumed equality on variance and for both alternative hypo
thesis
for (j in unique(tg_data$dose)) {
    for (test in c("greater", "less")) {
      for (variance in c(T,F)) {
        data_sub_set <- subset(tg_data, dose==j)
        temp <- t.test(len~supp, paired=F, var.equal=variance, data=data_sub_set, alter
native=test)
        res_sup_dos[k,"supplement1"] <- levels(data_sub_set$supp)[1]
        res_sup_dos[k,"supplement2"] <- levels(data_sub_set$supp)[2]
        res_sup_dos[k,"dose"] <- j
        res_sup_dos[k,"low_lim"] <- temp$conf.int[1]
        res_sup_dos[k,"up_lim"] <- temp$conf.int[2]
        res_sup_dos[k,"pval"] <- temp$p.value
        res_sup_dos[k,"variance_equal"] <- variance
        res_sup_dos[k,"test"] <- test
        k <- k+1
      }
    }
}
# P-Values with Bonferroni adjustment
res_sup_dos$adj_pval <- res_sup_dos$pval*50
# P-value interpretation
res_sup_dos$outcome <- ifelse(res_sup_dos$pval<=0.05, "Difference", "No Difference")
# Formating
res_sup_dos[c(4:6,9)] <- round(res_sup_dos[c(4:6,9)],3)
res_sup_dos_diff <- subset(res_sup_dos, outcome=="Difference")
res_sup_dos_diff <- res_sup_dos_diff[,c(1:6,9,7:8,10)]
kable(res_sup_dos_diff, format = "markdown")
```

	supplement1	supplement2	dose	low_lim	up_lim	pval	adj_pval	variance_equal	test	outcome
1	OJ	VC	0.5	2.378	Inf	0.003	0.133	TRUE	greater	Difference
2	OJ	VC	0.5	2.346	Inf	0.003	0.159	FALSE	greater	Difference
5	OJ	VC	1.0	3.380	Inf	0.000	0.020	TRUE	greater	Difference
6	OJ	VC	1.0	3.356	Inf	0.001	0.026	FALSE	greater	Difference

The table above shows the supplement comparisons for each dosage for which a statistical difference at 95% was observed. It shows that orange juice is more effective at 0.5 and 1 mg but also that both supplements are equally effective at 2mg. When we use the Bonferroni adjustment on the p-values, we can only conclude that orange juice is more effective at 1mg

- Which Dosage is Better?

```
# Define data frame to hold our analysis's results, use same columns as before
res_dos_bet = do.call(data.frame,
                      as.list(dose1=numeric(), dose2=numeric(), common_cols))
k=1
for (j in unique(tg_data$dose)) {
    for (test in c("greater","less")) {
      for (variance in c(T,F)) {
        data_sub_set <- subset(tg_data, dose!=j)
        temp <- t.test(len~dose, paired=F, var.equal=variance, data=data_sub_set, alter
native=test)
        res_dos_bet[k,"dose1"] <- unique(data_sub_set$dose)[1]
        res_dos_bet[k,"dose2"] <- unique(data_sub_set$dose)[2]
        res_dos_bet[k,"low_lim"] <- temp$conf.int[1]
        res_dos_bet[k,"up_lim"] <- temp$conf.int[2];
        res_dos_bet[k,"pval"] <- temp$p.value
        res_dos_bet[k,"variance_equal"] <- variance
        res_dos_bet[k,"test"] <- test
        k <- k+1
      }
    }
}
# P-Values with Bonferroni adjustment
res_dos_bet$adj_pval <- res_dos_bet$pval*50
# P-value interpretation
res_dos_bet$outcome <- ifelse(res_dos_bet$pval<=0.05, "Difference", "No Difference")
# Formating
res_dos_bet[c(3:5,8)] <- round(res_dos_bet[c(3:5,8)],3)
res_dos_bet$difference <- ifelse(res_dos_bet$outcome=="Difference", T, F)
res_dos_bet_diff <- subset(res_dos_bet, outcome=="Difference")
res_dos_bet_diff <- res_dos_bet_diff[c(1:5,8,6,7,9:10)]
kable(res_dos_bet_diff, format = "markdown")
```

	dose1	dose2	low_lim	up_lim	pval	adj_pval	variance_equal	test	outcome	difference
3	1.0	2	-Inf	-4.175	0	0	TRUE	less	Difference	TRUE
4	1.0	2	-Inf	-4.174	0	0	FALSE	less	Difference	TRUE
7	0.5	2	-Inf	-13.281	0	0	TRUE	less	Difference	TRUE
8	0.5	2	-Inf	-13.279	0	0	FALSE	less	Difference	TRUE
11	0.5	1	-Inf	-6.753	0	0	TRUE	less	Difference	TRUE
12	0.5	1	-Inf	-6.753	0	0	FALSE	less	Difference	TRUE

The table above shows the comparisons across supplements for each dosage for which a statistical difference at 95% was observed. It shows that 2mg is superior to 1 and 0.5 mgs, and that 1mg is superior to 0.5 mg. The results are valid even if we use a Bonferroni adjustment on the p-values

Conclusions

Overall after all the previous analysis, we conclude:

- The increase in dosage of either supplement will also increase the tooth length.
- The Type of supplement alone does not affect tooth growth.
- The supplement, orange juice (OJ), has greater impact to tooth growth than Vitamin C (VC) for dosage at 0.5mg and 1mg.
- However, when the dosage reaches 2mg, the impact by Orange Juice and Vitamin C is similar.

Assumptions

- The above conclusions assume that the data are not paired.
- The guinea pigs that receive different dose are totally separate. That means guinea pigs received dose 0.5 have nothing to do with dose 1.0, and dose 2.0